School Identification for Active Travel Interventions

Urban Analytics Group Project (URBAN 5124)
Chania Rahmah
Jorge Michel Gutierrez
Xinyue Wang
Cameron Wright
25th April 2022

1. Introduction

Sustrans is a charity whose primary aim is to encourage more people to use activities in the UK (Sustrans, 2022). In Scotland, they operate a variety of projects intended to increase the rate of pupils using active travel modes for their journey to and from school. Several interventions have been initiated to create an accessible and friendly route for active modes in Scotland, such as I Bike and Big Pedal, Bikeability and Cycling Friendly Schools (CFS), and WOW Travel Tracker (TT). These interventions are provided by the Active Travel Delivery Partners (ATDP), Cycling Scotland, Sustrans Scotland, and Living Streets. However, some schools remain to participate in interventions, by which hopefully, they would reduce their motorized mode of transport.

Sustrans has a need to identify which schools would benefit the most from future engagement and represent the greatest impact on their charitable aims. Using data collected by Sustrans on the circa 2,200 primary schools in Scotland, alongside data from other sources, we identify which schools have engaged the least with Sustrans recently and those schools likely to see the greatest benefits from future Sustrans interventions. To achieve this in an interactive format and to empower Sustrans staff to answer further questions, it was delivered an interactive ArcGIS Online dashboard designed to empower Sustrans staff to explore the data themselves.

Moreover, two questions shall be addressed: a) How does the frequency of road traffic incidents in the vicinity of a school correlate with the modal share of active travel for trips to and from that school? b) Are there common characteristics between the schools engaged the least with Sustrans' Hands Up Scotland Survey and their interventions?

Two statistical models were produced to address these questions: a multiple linear regression model and a logistic regression model, respectively. Additionally, a dashboard was created to facilitate further exploration and visualisation of the data by Sustrans.

2. Literature review

Sidharthan (2011) found that factors influencing travel to and from one New Zealand Aotearoa primary school included weather conditions, the distance between home and school and perceptions of safety. They also found that households located geographically proximate to one another tend to match one another's modal choice for school trips. This may be owing to sharing the same built environment, perceived safety levels and distance from school but social factors may partly explain it: if one household's children walk to school, greater legitimacy is granted to this mode for other nearby families who interact with the former.

Ewing (2004) additionally found that in a US environment, the size of the school was associated with a higher rate of driving to school but only insofar as the higher pupil roll also implied a wider catchment area. It is not expected that this effect transfers to a higher pupil roll in a densely populated catchment area such as in the centres of Scottish cities.

Surrounding land uses of school sites have been observed to have less effect on modal choice for school trips than for travel in general. It is suggested that this may be owing to a tendency for school trips to be made in isolation rather than chained together with trips for other purposes. A positive correlation between road traffic incident frequency and socio-economic deprivation has been observed, and high lack is associated with high attendance rates at hospital accident and emergency (A&E) departments. (S. Boniface, 2015)

Regarding the selection of the tool for producing the dashboard, the ESRI platform allows efficient ways to interact with spatial data and its operational dashboards were optimal for the purposes of the project as they suit descriptive statistics and work smoothly for describing events that have occurred as opposed to predicting events.(Miller, 2021).

3. Data and Feature Engineering

The various dataset is required in this school identification project. The detail of the dataset and the sources could be found in the table below:

Table 1. Data and Sources

Data	Dates	Source
School locations mapped in GIS	2020	Scottish Government
Child KSI/road traffic accidents from STATS19	2015-2020	UK Department for Transport
Pupil Equity funding (PEF)	2019-2020	Scottish Government
School intervention data (Big Pedal, TT, I Bike, Bikeability)	2017-2020	Sustrans
Hands Up Scotland Survey data – school level	2015-2020	Sustrans
Scottish Index of Multiple Deprivation (SIMD)	2021	Scottish Government
School roll numbers – number of pupils in each school	2015-2020	Scottish Government

3.1 Road Safety and Casualty data

As an indicator of how safe is the road network around the schools for the students to commute by active modes of transport, it is considered the number of accidents that have occurred from 2016 to May 2021, which is the most recent published data by the UK Department of Transport. The accident data is made available in 3 separate datasets: 'accidents', 'casualties', and 'vehicles' which were linked together by a unique accident ID and included information about the location, date and time, accident causes, details of people and vehicles involved, weather conditions, etc.

After merging the datasets for 2016-2021, the data was filtered to contain only casualties involving persons younger than 16 years old, that were either cycling or walking and that occurred in any Scottish local authority. The resulting dataset contained only 2,491 observations corresponding to each person injured in any accidents.

The relevant accidents for this research are the ones occurring nearby the schools, so for this, the coordinate data was used to locate the casualties, as well as the schools. The radius considered for the accident 'catchment area' of each school is 1.5 miles, which corresponds to the average travel length for primary schools in England (Sustrans, 2018). A spatial join method, from the package in R, was used to aggregate casualty data and assign the count of casualties within the threshold to each school. The outcome is a spatial dataset containing the original school information but with an attribute of a number of casualties. Important to note that there is particular interest from Sustrans in KSI (Kills or serious accidents), so a count for accidents in this category is also calculated with the same methods.

3.2 Pupil Equity Funding

Pupil Equity Funding (PEF) is a form of financial support granted by the Scottish Government to support children who experience barriers to learning because their family is experiencing poverty or financial hardship. This funding is provided to each school attended by eligible children. The schools receive an additional £1,200 per year for every child registered as a pupil in financial difficulties (Scottish Government, 2018). For the purposes of our research, PEF is treated as a proxy for the poverty level at each school.

In this data processing, we utilise the PEF data from the year 2020. In this data processing step, the PEF in each school is categorised into 10 quintiles, as stated in the table below:

Table 2. PEF Quintiles

Quintile	0	1	2	3	4	5	6	7	8	9	10
Values	0	1200	6000	12000	19440	30000	42120	58320	78000	108480	290400

The funding minimum allocation is assumed to be around £1,200 and £6,000. Meanwhile, the highest funding is as much as £290,400.

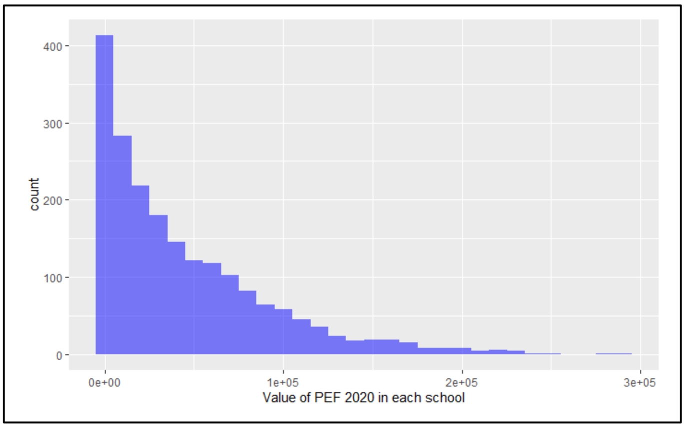


Figure 1. Value of PEF in each School

The histogram plot above illustrates the distribution of PEF allocation in each school. In 2020 there are 1,900 primary schools that receive PEF. Most schools received PEF allocation under £10,000 while the rest small number of the schools accepted more than that.

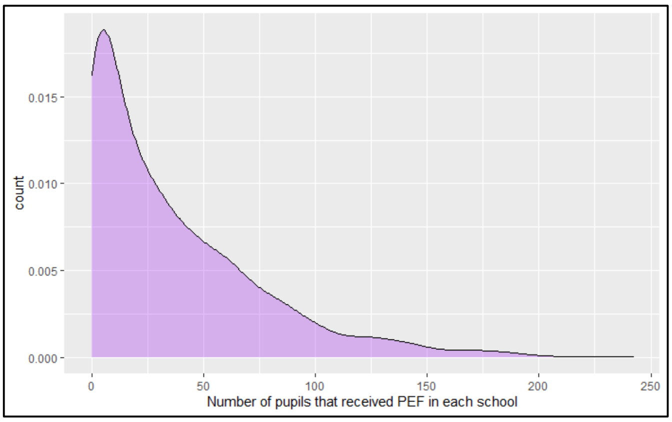


Figure 2. Number of Pupils that Received PEF in each School

From the PEF data, we could further examine the number of pupils who registered as a pupil in need and require funding support. The density plot above represents the number of pupils that received PEF in each school. It is ranging from 7 to 242 pupils have been registered for financial support from PEF. Furthermore, in this data processing, we established a map to better understand the distribution of 20% of schools that receive the highest allocation of PEF.

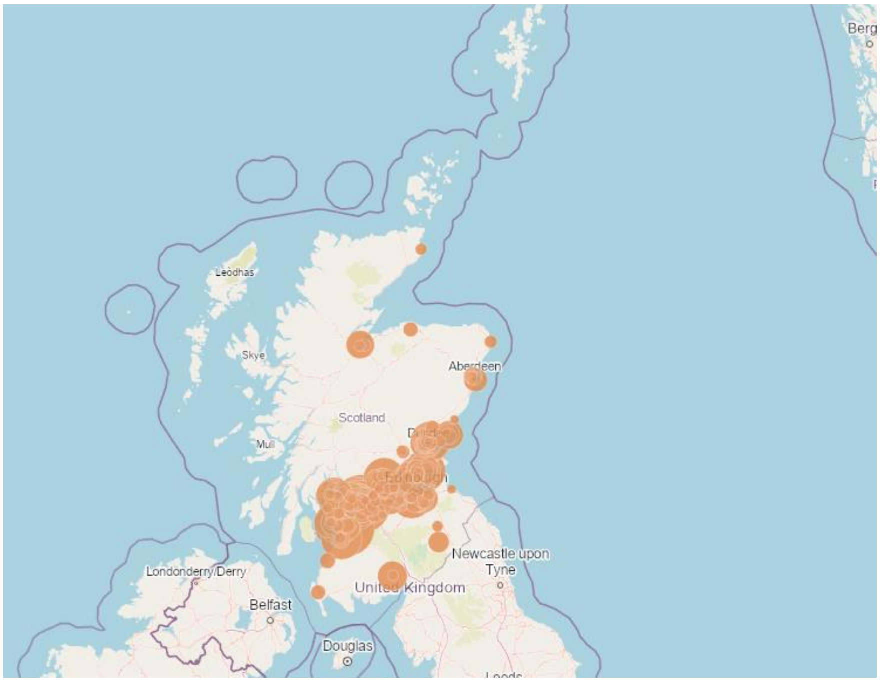


Figure 3. Map of the 20% Highest PEF Distribution

From the visualisation map, we could see most of the distribution of the highest receivers of PEF schools located in the urban area, such as Edinburgh and Glasgow.

3.3 Intervention data

1. Big Pedal

Sustrans' Big Pedal (now known as the Big Walk and Wheel) is a 10-day school intervention organised annually to motivate schools to encourage households to use active travel to get to and from school. Participating schools compete to win prizes. Schools that record 15% or more of their role as taking an active trip on a single day are entered into a prize draw to receive equipment intended to further enable active travel.

Data on participating schools was provided by Sustrans for three consecutive academic years: 2016-17, 2017-18 and 2018-19. Although various data were provided about the participating schools, only each school's participation in the programme each year was deemed relevant to this project. As with other datasets, schools were uniquely identified by their SEED code and only appeared in each year's data if they participated.

Each year's data required some amount of manual cleaning before it could be usefully machine-read. This included removing cover sheets from the provided Excel spreadsheets and extraneous header rows from the main datasheets. For simplicity, the cleaned spreadsheets were saved in comma-separated values (CSV) format. Once loaded into R, the three years were merged into a single dataset with a row for each participating school and participation in each year encoded as a column for each year and a value of either 1 or 0 for each school in each year representing participation or not respectively.

2. Bikeability

Sustrans provide us the data on schools that have delivered Bike ability training from our partners at Cycling Scotland. The data we report on includes all the Primary Schools and some of secondary schools from 2016 to 2020 which was divided into 4 sheets. Because of, mainly smaller rural schools, delivering once every 2 or 3 years, the annual data in any one year doesn't translate directly to the number of schools that are actively participating in the programme. In addition to these, as a result of school closures during the 2019/20 and 2020/21 academic sessions, and associated public health restrictions, delivery of Bikeability Scotland was negatively impacted. For simplicity, the spreadsheets after manual cleaning were saved in comma-separated values (CSV) format.

The data columns include three types of schools, and they are primary schools, secondary schools and secondary only. Given that the dashboard will focus on the primary schools, so we filtered all the data by primary schools and merged the data set in different years by their unique SEED code after loading them into R. It is worth noting that the different level of engagements on Bikeability was remarked as level 1 and level 2. And then the engagement was encoded as a new column for each year and a value of either 1 or 0 for each school in each year representing at least one level or not even once respectively.

3.4 Hands up Scotland Survey

The Hands up Scotland Survey (HUSS) is a survey designed to provide up to date information on school pupils' mode of travel to school. The dataset contains the school names and the information about the number of pupils in each transport. This survey is conducted annually by Sustrans and Scottish local authorities and funded by Transport Scotland. In this analysis, we will utilize HUSS from the years 2019 and 2020. The goals of using this dataset are to identify schools that have done and have not done HUSS in those two years.

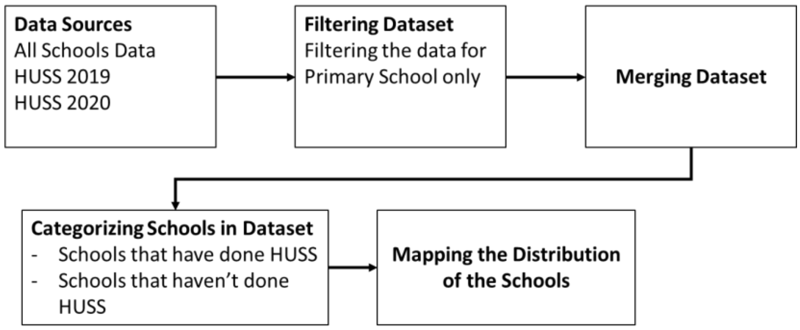


Figure 4. HUSS Data Processing

The flowchart above illustrates the data processing flows. The first step of this process is to filter the HUSS information for primary schools only. The next step is to merge the two years dataset, the result is the list of school names that have done HUSS in either 2019, 2020, or both years.

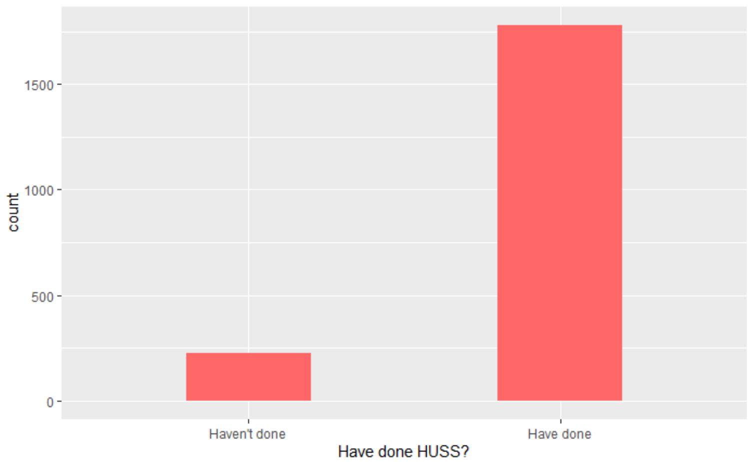


Figure 5. HUSS Primary School Engagement

By merging the 2003 primary school names list and the school names that have been done HUSS, we could group the school into two categories. The categories are: 224 schools that have not done HUSS, and 1779 have done HUSS.

After knowing the category of school and the HUSS engagement, we plot the distribution on the map. The first map below shows the distribution of primary schools that have been done HUSS. Most of the schools that have done HUSS exist in the urban area such as Aberdeen, Dundee, Edinburgh, and Glasgow.

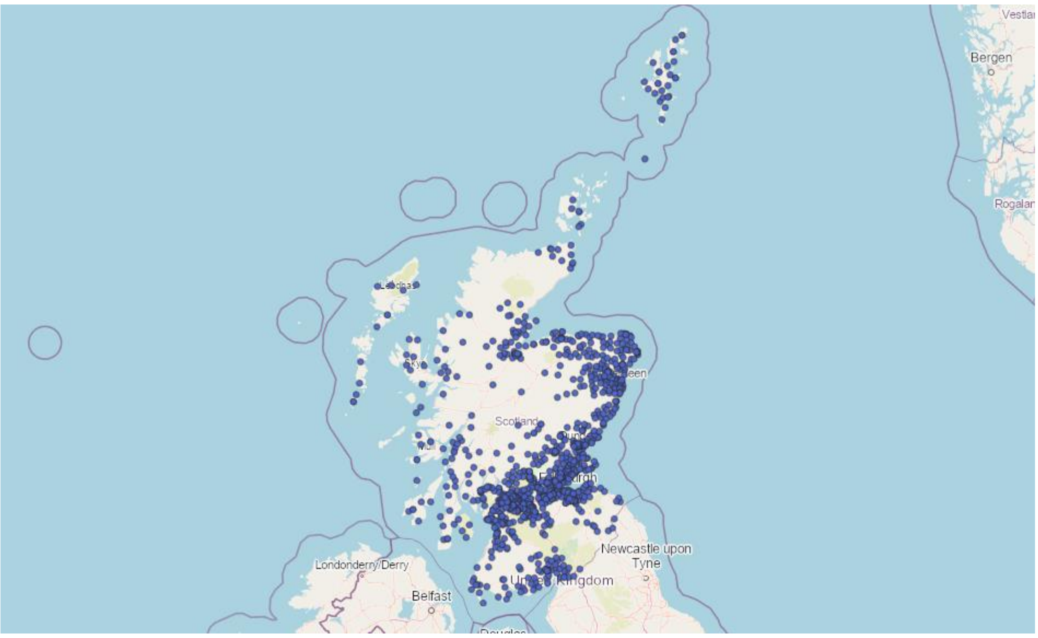


Figure 6. Distribution of Schools that Have Done HUSS

Furthermore, we also create a map for the other school category. Even though there are some schools in the urban area especially Glasgow that haven't done HUSS, there is a tendency that schools that have not done HUSS to be located in the rural area or in the island area around Scotland. The distribution of schools that have not done HUSS could be found on the map below:

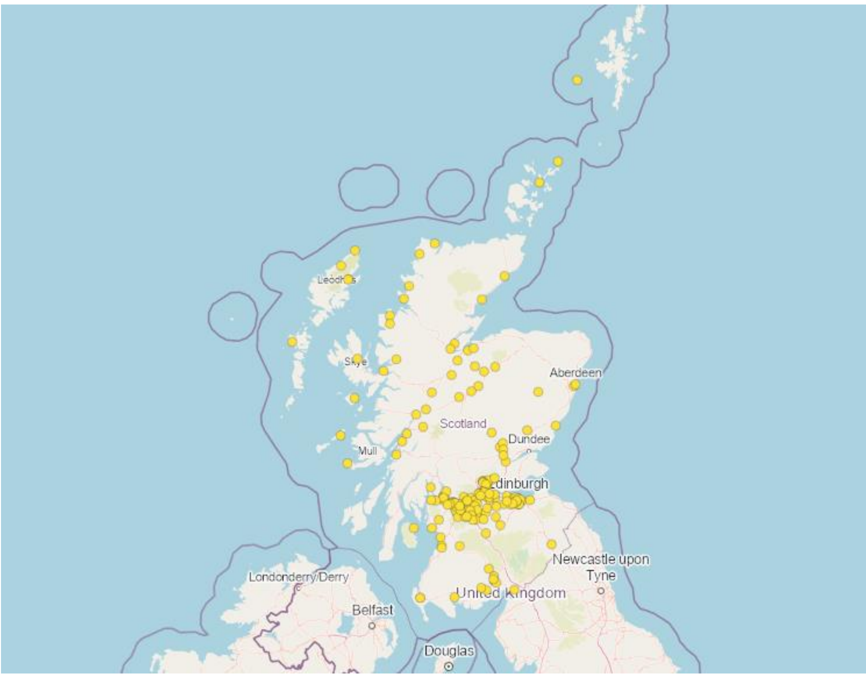


Figure 7. Distribution of Schools that Have Not Done HUSS

The Scottish government provides statistical data at a school level for every school in the country which was used to add relevant data about the schools. Two main variables were extracted from the 2021 dataset: the number of pupils and the Urban-Rural Classification. The latter is aligned with the government classification developed to improve the understanding of the different issues that each urbanization category is facing.

From the possible 8 possible classifications, it was found that all schools are within the 6 most accessible. The following table details the criteria for these classifications: (Scottish Government, 2021):

Table 3. Urban-Rural Classification

Urban-Rural Classification	Definition
Large Urban Areas	settlements of 125,000 people and over
Other Urban Areas	settlements of 10,000 to 124,999 people
Accessible Small Towns	settlements of 3,000 to 9,999 people, and within a 30-minute drive time of a settlement of 10,000 or more
Remote Small Towns	settlements of 3,000 to 9,999 people, and with a drive time of over 30 minutes but less than or equal to 60 minutes to a settlement of 10,000 or more
Accessible Rural Areas	areas with a population of less than 3,000 people, and within a drive time of 30 minutes to a settlement of 10,000 or more
Remote Rural Areas	areas with a population of less than 3,000 people, and with a drive time of over 30 minutes but less than or equal to 60 minutes to a settlement of 10,000 or more

4. Methods

4.1 Merging all data

All the data that have been through the processing step were merged into the same file and format. The final dataset contains 28 variables that are explained in the table below:

Table 4. Final Dataset Variables

No	Variable Name	Definition
1	SEED	SEED code for school
2	SchoolName	School name
3	LAName	Local authority name
4	HUSS_Survey_Done	HUSS survey engagement
5	HUSS1920_cycling	The average percentage of pupils cycling in HUSS 2019 and 2020
6	HUSS1920_walking	The average percentage of pupils walking in HUSS 2019 and 2020

No	Variable Name	Definition
7	HUSS1920_scooter	The average percentage of pupils travelling to school by scooter in HUSS 2019 and 2020
8	HUSS1920_parkstride	The average percentage of pupils travelling to school by park and stride in HUSS 2019 and 2020
9	HUSS1920_driving	The average percentage of pupils driving in HUSS 2019 and 2020
10	HUSS1920_bus	The average percentage of pupils travelling to school by bus in HUSS 2019 and 2020
11	HUSS1920_taxi	The average percentage of pupils travelling to school by taxi in HUSS 2019 and 2020
12	HUSS1920_other	The average percentage of pupils travelling to school by another transport mode in HUSS 2019 and 2020
13	HUSS1920_ActiveTravel	Sum of the active travel transport mode percentage
14	HUSS1920_Motorized	Sum of the motorized transport mode percentage
15	PEF_Allocation_Before_Top_Up	PEF allocation before top-up
16	PEF_Top_Up	PEF top-up
17	PEF_Allocation_After_Top_Up	Total PEF allocation after top-up
18	SIMD_Quintile	SIMD quintile categories
19	Latitude	Geographic information
20	Longitude	Geographic information
21	Casualties_KSI	The number of Kill, Serious, or Injuries road accident Casualties within 1.5 miles radius of the school
22	Casualties_total	The number of total road accident Casualties within 1.5 miles radius of the school
23	Pupil.roll	Pupils roll in each school
24	Urban.Rural.Class	Urban-rural classification
25	Intervent	Intervention engagement
26	PEF_decile	PEF value decile
27	PEF_pupils	Number of pupils that received PEF
28	PEF_percentage	Percentage of pupils that received PEF

In the data merging process, we found there is a missing value issue in some of the variables. In HUSS data, transport mode information is having missing values because some of the schools didn't engage in HUSS. The other missing values also exist in school information such as pupil roll and urban-rural classification of the school. Therefore, we encode the missing value to be -1 in order to make it readable for the dashboard because the dashboard could not read the data with missing values.

4.2 Dashboard construction

The main objective for this practical project was to produce a dashboard where the organization Sustrans is able to visualize and target specific schools that would maximize the impact of active travel interventions. Among the options to create data dashboards, the ArcGIS online platform was chosen mainly due to its compatibility with the current tools being used by some Sustrans' teams. ArcGIS online platform allows creating dashboards based on maps created on any of the platform's GIS software, providing the ability to manipulate and visualize data in an interactive way.

The design process of the dashboard consisted first in linking the main question and research aims to the dashboard functionalities so it is as useful as possible. A mix of input, output and visualization components was chosen to achieve the objectives of the research.

Input

To answer the question *Which schools in Scotland have the greatest potential for change/increasing active modes (walking, scootering and cycling)?* it is necessary to look at different variables such as HUSS and previous interventions engagement, economic indicators or accident rates. Thus, filters and data selectors were the most suitable to allow the user to query information that then would be displayed. The following table details the functionality of each of the 8 filters built into the dashboard:

Table 5. The Functionality of Data Filter

No	Filter	Type	Description
1.	HUSS participation	Dropdown list with 2 possible values	Queries schools that have participated in answering HUSS either in 2019 or 2020
2.	Active travel interventions	Dropdown list with 2 possibles values	Queries schools that have participated in any intervention
3.	KSI casualties	Number range slider selector	Filters schools according to the amount of serious or fatal casualties that have occurred in a 1.5-mile radius from 2016 to May 2021.
4.	Area Classification	Dropdown list with 6 values	Filters schools according to the Scottish rural-urban classification.
5.	SIMD	Number range spinner selector	Queries schools by SIMD Quintiles
6.	PEF	Number range slider selector	Filters schools by the percentage of students that received funds. Missing data coded as -1
7.	Driving rates	Number range slider selector	Filters schools that have done HUSS according to the share of students travelling to school by car.
8.	Active Travel	Number range slider selector	Filters schools that have done HUSS according to the share of students travelling to school by scooter, walking or cycling.

Visualization

The main visualization component is a map located in the middle of the dashboard, which contains two layers: first, a school layer, for which the source is a dataset containing all the information about each school, the same one that is used for all the dashboard features. The second layer is the location of road accidents involving people less than 16 years old and active modes of transport. This map allows seeing detailed information about each casualty such as vehicles involved, date, demographics, etc. The displayed schools on the map are according to the selections made with the input features explained before. In addition, it is possible to see the most important details of a school by clicking on it on the map, for instance, the name, seed code, local authority, pupil roll, slight and serious accidents around it, etc.

Besides the map, there is a set of charts and indicators that not only shows data in a visual way but can contribute to gaining insights from the dashboard when they are used in combination with the input futures. The visualization includes 2 pies charts displaying the proportion of schools that have engaged in HUSS and interventions, one bar chart for the number of schools per SIMD Quintile, one gauge type indicator of the average percentage of active travel among the filtered schools, one numeric indicator on top to show how many schools are being looked at while manipulating the data. Finally, when clicking on a specific school on the map, a pie chart shows the total share of transport modes, if the data exists for that school.

Outcome

The purpose of the dashboard is to turn diverse data into actionable insights, therefore, it is very useful to be able to extract and download the information produced and save it in a format that allows sharing and further manipulation. To achieve this, we created a table that not only shows the seed code and school names of the queried schools, but it can be downloaded into a CSV once the desired criteria are applied to the filters.

4.3 Schools that don’t do HUSS and don’t engage in interventions, are there any common characteristics?

To answer this question, we will perform correlation analysis and statistical characteristic analysis on variables. Due to the large number of variables contained in the dataset, the regression model was created to filter out the independent variables that had a significant impact on the dependent variables, that is, to screen out the significant factors that affected the school's participation in the HUSS Survey and the school's participation in various intervention activities (including Bicycle, Scooter Parking and Ibike), and then further analysed and compared the variable characteristics.

In the previous data processing, when studying whether the school had participated in the Hands Up Scotland Survey or any School intervention, the dependent variable was encoded, 1 indicating that it had participated once or more, and 0 indicating that it had not participated at all. Since the variables are shown as binary, we chose the logistic regression in the following study.

The final processed dataset contains a total of 23 variables that could have a significant effect on the independent variables. Before studying the influence relationship, it is first necessary to have a different relationship between the independent variables X and Y before it can further have an influence relationship. To simplify the model and improve the fit, chi-square and ANOVA are first used to tentatively understand the differences between each influencing factor X and Y. After completing the difference analysis, we put X into the model for binary logistics regression.

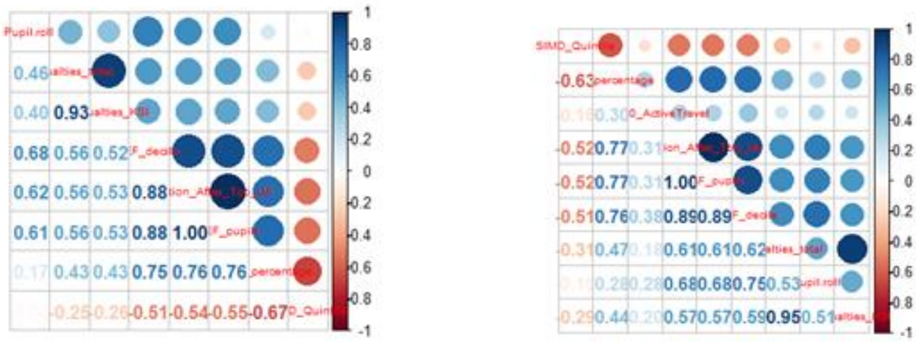


Figure 8. Correlation Between HUSS and Other Values

After logistics regression, we also generated diagnostic maps to test how well the model fits. Finally, the statistical characteristics of the two sets of data and the p-value of the model are compared and analysed.

4.4 Do schools with accidents nearby have higher walking, scooting and cycling figures in HUSS?

To answer this question, we will establish a Pearson correlation analysis and create a multiple linear regression model to find the correlation between accident rates and transport modes to school such as driving, cycling, scooting, or even walking.

Pearson correlation analysis is a bivariate analysis that produces a correlation coefficient. The coefficient could measure the strength and direction of the linear relationship between two variables that help to describe the direction and degree of the correlation (Bolboaca and Jantschi, 2006). In this analysis, we will try to see the relation between accident rates and each of the transport modes that pupils use to go to school.

Furthermore, we also try to find the causal relationship between accident rates and transport modes. In this case, we utilize multiple linear regression with accident rate as the dependent variable. However, we found that the accident rates variable is not distributed normally thus we applied a log transformation to it.

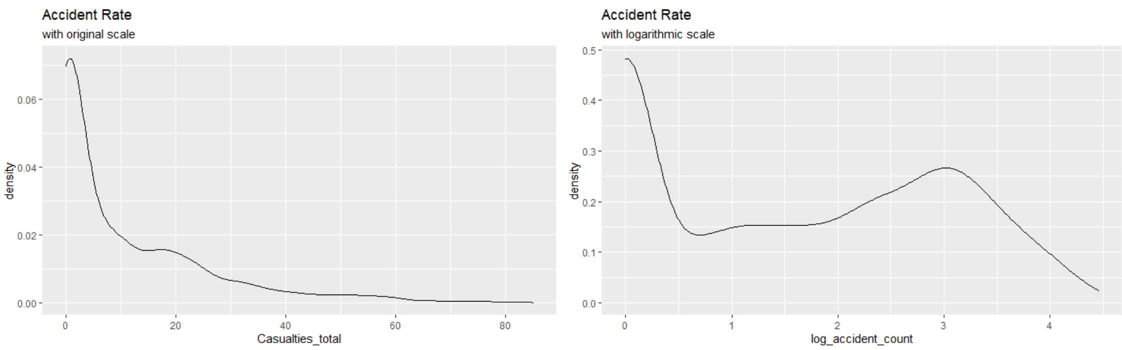


Figure 9. Normal Distribution of The Dependent Variable

5. Results

5.1 Results and discussion for question 1

1. Binary Logistic Regression

Initially, a binary Logistic regression analysis was performed directly on a variable, and an overfitting problem occurred. Without being able to increase the data set, we remodeled the independent variables grouped to test the effect of the independent variables on the dependent variables. As variables related travel modes of these schools are completely decided by the results of HUSS Survey Done, so we drop the relevant variables when we try to make a regression model.

In fact, according to the chi-square independence test, the p-values of the variables related to Active travel are greater than the significance level (set to 0.05).

```
        Pearson's Chi-squared test

data:  test_travel
X-squared = 1986, df = 1657, p-value = 3.587e-08
```

After the test, we dropped the active travel variables and created the models. Then we get two groups of binary regression models as the following tables show.

Logistics Regression Models.

	(1) HUSS	(2) HUSS	(3) HUSS	(4) HUSS	(5) intervent	(6) intervent	(7) intervent	(8) intervent
(Intercept)	1.543***	2.172***	2.258***	2.481***	0.483***	0.392*	0.350	0.858***
	(0.170)	(0.278)	(0.287)	(0.370)	(0.123)	(0.188)	(0.196)	(0.257)
SIMD_Quintile	0.206***	0.143*	0.081	0.043	0.174***	0.231***	0.167***	0.058
	(0.059)	(0.060)	(0.064)	(0.079)	(0.041)	(0.045)	(0.048)	(0.058)
Urban.Rural.ClassAccessible small towns		–0.555	–0.734*	–0.690*		0.297	–0.056	0.009
		(0.320)	(0.338)	(0.343)		(0.232)	(0.246)	(0.249)
Urban.Rural.ClassLarge urban areas		–0.842***	–0.570	–0.516		0.329	0.202	0.303
		(0.246)	(0.365)	(0.371)		(0.173)	(0.276)	(0.281)
Urban.Rural.ClassOther urban areas		–0.505*	–0.534	–0.487		0.019	–0.269	–0.190
		(0.241)	(0.282)	(0.288)		(0.154)	(0.190)	(0.195)
Urban.Rural.ClassRemote rural areas		–0.404	–0.366	–0.387		–0.626***	–0.525**	–0.562***
		(0.265)	(0.267)	(0.268)		(0.159)	(0.161)	(0.162)
Urban.Rural.ClassRemote small towns		1.485	1.296	1.354		–0.953***	–1.351***	–1.246***
		(1.028)	(1.033)	(1.035)		(0.285)	(0.301)	(0.304)
Pupil.roll			0.001	0.001			0.003***	0.003***
			(0.001)	(0.001)			(0.001)	(0.001)
Casualties_total			–0.019**	–0.018*			–0.015*	–0.011
			(0.007)	(0.007)			(0.006)	(0.006)
PEF_Allocation_After_Top_Up				0.000				–0.000
				(0.000)				(0.000)
PEF_percentage				–0.008				–0.013
				(0.011)				(0.007)
McFadden's R^2	0.009	0.024	0.031	0.032	0.008	0.029	0.042	0.047
Nagelkerke's R^2	0.013	0.033	0.043	0.044	0.013	0.049	0.069	0.077
AIC	1353.297	1343.092	1337.173	1340.239	2321.473	2281.603	2256.942	2249.061
BIC	1364.485	1382.249	1387.518	1401.772	2332.661	2320.760	2307.287	2310.593
Num. obs.	1986	1986	1986	1986	1986	1986	1986	1986

Note. Unstandardized regression coefficients are displayed, with standard errors in parentheses.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Figure 10. The Stepwise Binary Regression Models

The model 1 of HUSS table indicates that the SIMD quintile is positive correlated with the participation of HUSS Survey, which is the same with the trend of the coefficient between SIMD and the engagement of intervention activities. In addition, the variable is significant to both dependent variables in the first group of models.

In the model 2, we added the variables related to regions where the schools belong to. Compared with the above tables, we can find that the influence of SIMD on the participation of HUSS Survey reduced which is still significant to the engagement of intervention activities. And the significance of the location variables to the two dependent variables is totally different. In the urban areas, the huss participations are more related to by the location factors while the intervention engagements are more affected by the rural areas or the remote towns.

Then we added pupil rolls and total casualties and number of pupils into the model 3, we can see that the number of pupils is insignificant to HUSS while both the engagement of intervention activities and participation of HUSS are negatively correlated with the variable of casualties. At the last we added the variables related to PEF which is also insignificant.

In the above process, we can find that there are many variables that have changed in different models. So, considering excluding these insignificant variables, we used the stepwise regression method here to select variables. Fortunately, the filtered variables of the optimization are the same. And the final models are these and the R2 of the models are the top among all the models which means the degrees of fitting are the highest.

	Intervent					HUSS Survey				
	Estimate	**S.E.**	**z**	**p**		**Estimate**	**S.E.**	**z**	**p**	
(Intercept)	1.133	(0.138)	8.195	< .001	***	2.624	(0.223)	11.771	< .001	***
Urban.Rural.ClassAccessible small towns	0.007	(0.246)	0.026	.979		−0.712	(0.338)	−2.104	.035	*
Urban.Rural.ClassLarge urban areas	0.302	(0.280)	1.079	.280		−0.532	(0.368)	−1.444	.149	
Urban.Rural.ClassOther urban areas	−0.203	(0.191)	−1.062	.288		−0.512	(0.283)	−1.812	.070	.
Urban.Rural.ClassRemote rural areas	−0.585	(0.161)	−3.632	< .001	***	−0.393	(0.267)	−1.472	.141	
Urban.Rural.ClassRemote small towns	−1.238	(0.303)	−4.081	< .001	***	1.342	(1.034)	1.298	.194	
Pupil.roll	0.003	(0.001)	4.579	< .001	***	0.001	(0.001)	1.809	.070	.
Casualties_total	−0.012	(0.006)	−1.934	.053	.	−0.018	(0.007)	−2.512	.012	*
PEF_percentage	−0.020	(0.004)	−4.735	< .001	***	−0.009	(0.006)	−1.508	.131	
McFadden's R^2	0.032					0.046				
Nagelkerke's R^2	0.043					0.076				
AIC	1336.533					2246.894				
BIC	1386.878					2297.238				
Num. obs.	1986					1986				

Note. Unstandardized regression coefficients are displayed, with standard errors in parentheses.
* $p < .05$. ** $p < .01$. *** $p < .001$.

Figure 11. Comparison of Final Regression Models

According to the comparison of final regression models, we can find that both the engagement of intervention activities and participation of HUSS are negatively related to the Other urban areas, remote rural areas, total casualties and PEF percentage. In the next section, we filter the data based on these variables and further analyse the distribution characteristics of variables.

2. Distribution characteristics

Next, we get the distribution characteristics of the selected schools which are not engaged in intervention or never attend the huss survey.

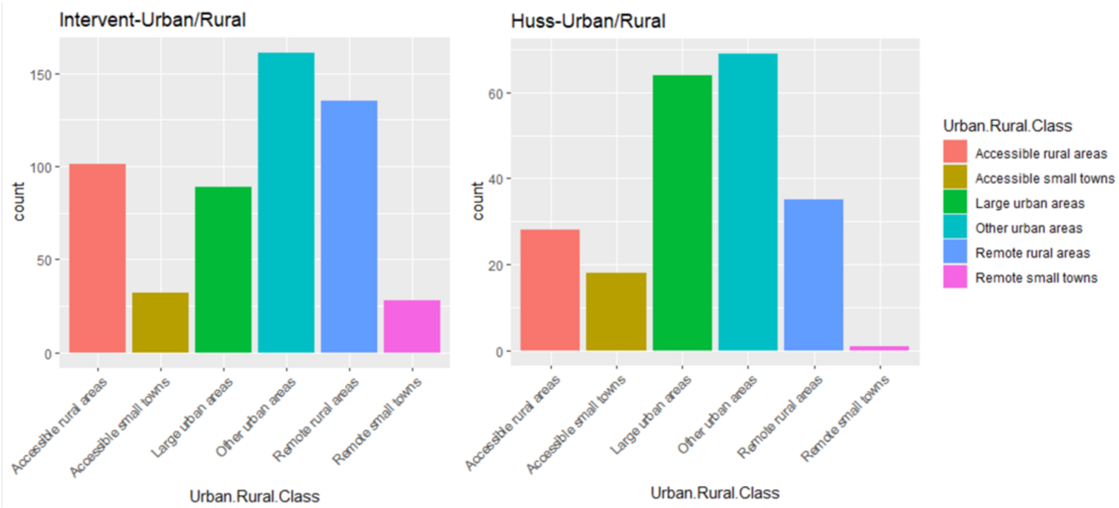


Figure 12. Characteristics of Location Distribution

The first comparison is about the location distribution. As most the target schools are in other urban areas and there are fewest ones in the remote towns. It is probably because the source of the data is more from the urban areas rather than the remote towns.

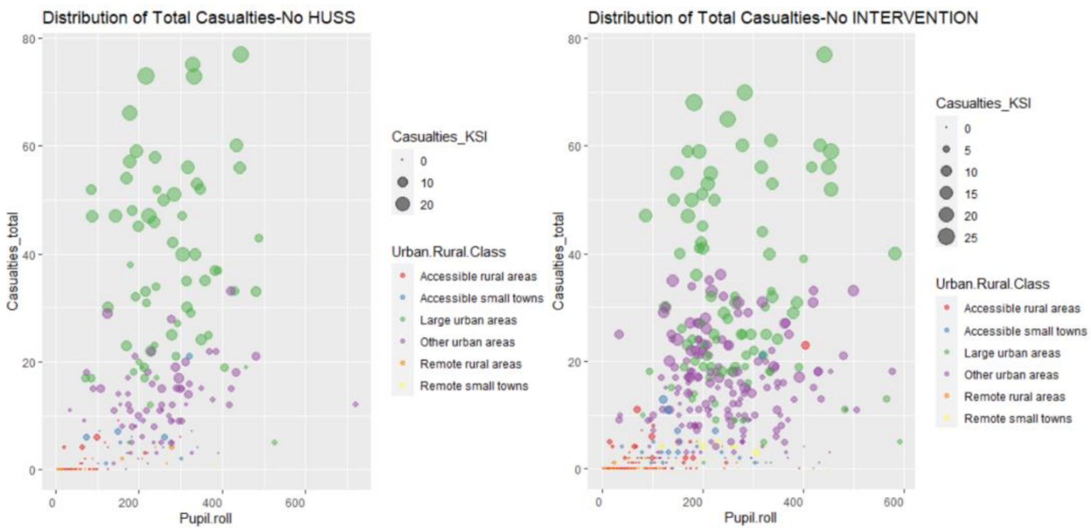


Figure 13. Characteristics of Total Casualties

And according to the further analysis contrast, the number of minors in these schools is mostly in the range of 200 to 400. At the same time, the maximum number of school accidents in both sets of data is less than 80. In addition, according to the accident situation distribution map, both sets of data show that accidents occur more often in urban areas. And there is no doubt that the KSI (kills or serious injuries) in these areas are also more severe.

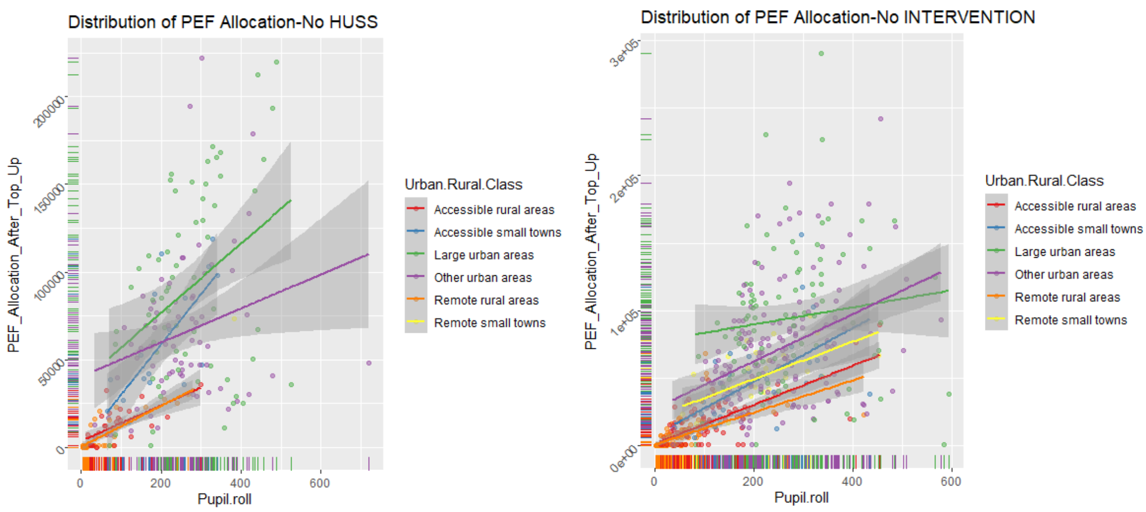


Figure 14. Characteristics of PEF Allocation

According to the distribution map of PEF Allocation, the common distribution characteristics of the two sets of data have three points.

First, accessible rural areas and remote rural areas have the fewest number of minors in schools, and schools with larger numbers of minors are concentrated in large urban areas and other urban areas. This feature is particularly pronounced in the No-HUSS group. The reason may be PEF was allocated by the number of pupils in different schools.

Second, the line segments in the figure above fit the classical linear regression model, and overall, the number of minors in each school and the number of funding assigned to them are positively correlated.

Third, the schools in rural areas get fewer allocations than the schools in urban areas. It means the PEF could pay more attention to the rural areas.

5.2 Results and discussion for question: Do schools with accidents nearby have higher walking, scooting and cycling figures in HUSS?

To find the correlation between transport mode and accident rate, this study utilizes correlation analysis. The correlation performs by the Pearson correlation method.

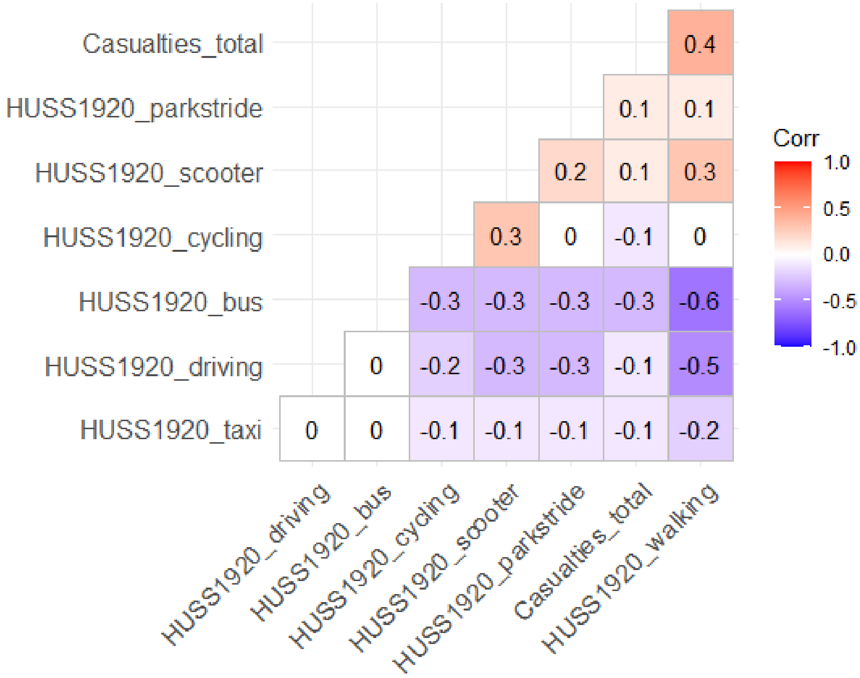


Figure 15. Correlation Matrix of Accident Rate and Transport Mode

The correlation matrix above shows the correlation between the accident rate and the transport modes. The accident rate is explained by casualties' data that was achieved from the calculation of children casualties from the accident that happens within a radius of 1.5 miles from the schools.

The analysis shows that the number of casualties has a correlation with transport modes both positively and negatively. The highest correlation exists between the number of casualties and walking, it indicates a 0.4 positive correlation. While park and stride also scooter, modes have a 0.1 positive correlation to the accident rate. On the other hand, cycling, driving, taxi, and bus are negatively correlated to accident rates as much as -0.1, and –03 for bus mode. The positive correlation here indicates that the two variables have a linear correlation which means as those transport mode users rise the number of accidents will arise too. Meanwhile, the negative correlation means that as those transport mode users rise the number of accidents conversely declines.

To better see the correlation between accident rate and each transport mode, we will create scatter plots. If the scatter plot shows a perfect linear correlation or a positive correlation, the distribution of the sample that is represented by points will form a diagonal line. In this case, we could see that the relation between accident rate and those four transport modes slightly presents a diagonal line.

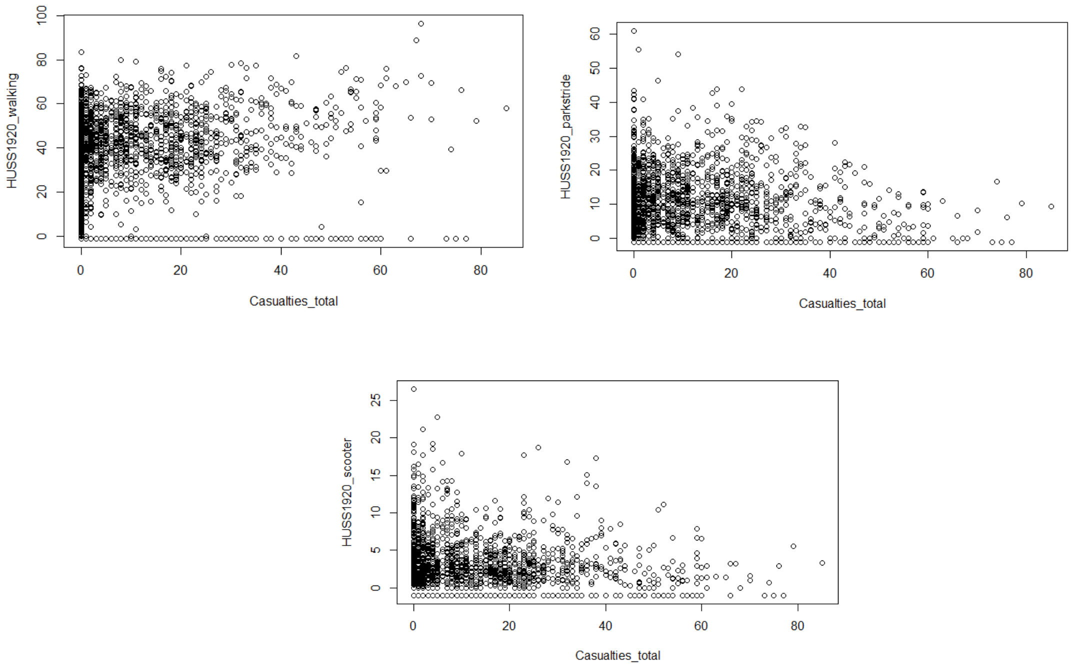
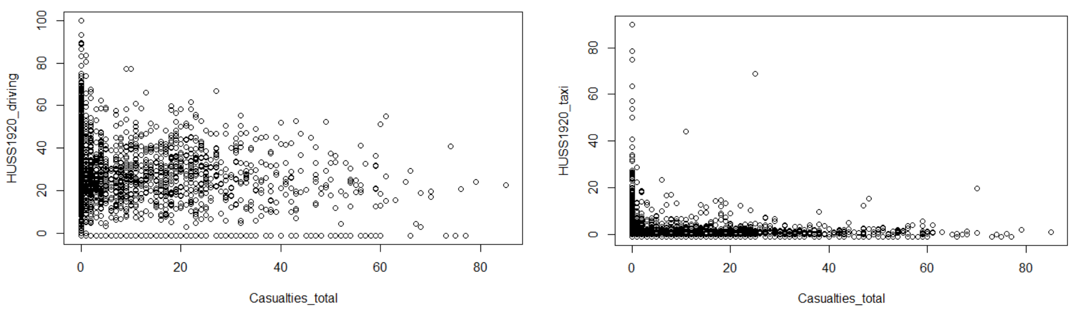


Figure 16. Scatter Plot of Positive Correlation

In the other side, the scatter plots below show the negative correlation between the accident rate and the transport mode.



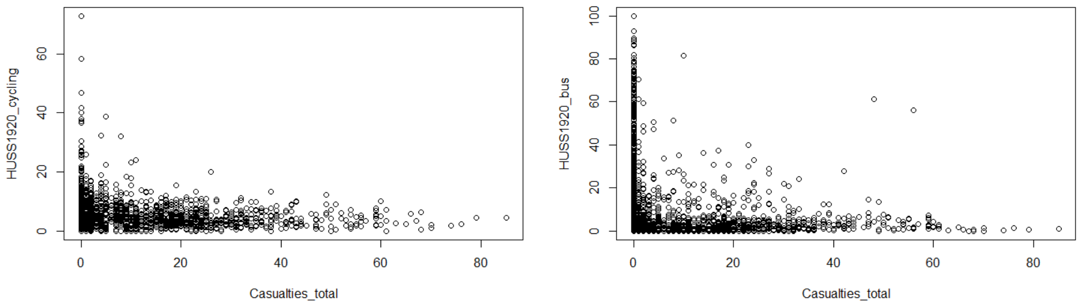


Figure 17. Scatter Plot of Negative Correlation

The correlation analysis could only provide a bivariate analysis or a correlation between two variables. Furthermore, we undertake multiple linear regression to understand the causal relationship between accident rates, transport modes, and other influential variables. In this analysis, we categorize the transport modes into two categories namely active travel and motorized. We quantified the value based on the percentage of the pupil who uses the active travel mode or the motorized one. Moreover, we added some other variables that might affect the accident rates such as the urban-rural class of the school, school size (pupil roll), deprivation level (SIMD_quintile), and the active travel intervention.

Table 6. Regression Model

The Relationship between Accident Rate and Transport Modes

	Dependent variable:	
	Casualties_totallog_accident	
	Model 1	
	(1)	(2)
Constant	5.60	0.70
	(1.31)***	(0.10)***
HUSS1920_ActiveTravel	0.02	0.002
	(0.01)	(0.001)*
HUSS1920_Motorized	0.02	0.002
	(0.02)	(0.001)
Urban.Rural.ClassAccessible small towns	-0.15	0.55
	(0.90)	(0.07)***
Urban.Rural.ClassLarge urban areas	25.31	2.50
	(0.78)***	(0.06)***
Urban.Rural.ClassOther urban areas	9.36	1.75
	(0.67)***	(0.05)***
Urban.Rural.ClassRemote rural areas	-0.82	-0.23
	(0.66)	(0.05)***
Urban.Rural.ClassRemote small towns	-0.55	0.55
	(1.23)	(0.09)***
Pupil.roll	0.01	0.001
	(0.002)***	(0.0001)***
SIMD_Quintile	-1.83	-0.16
	(0.18)***	(0.01)***
interventionWith Intervention	-0.95	-0.09
	(0.46)**	(0.03)***
Observations	1,778	1,778
R^2	0.64	0.80
Adjusted R^2	0.64	0.80
Residual Std. Error (df = 1767)	8.48	0.62
Note:	*p<0.1;* ***p<0.05;*** p<0.01	

Source: Hlavac, 2018

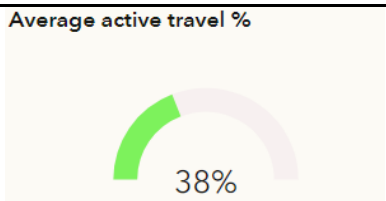
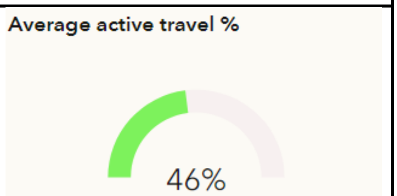
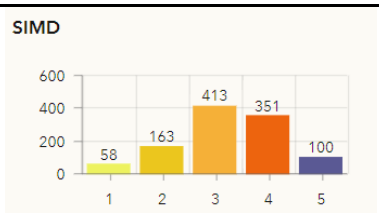
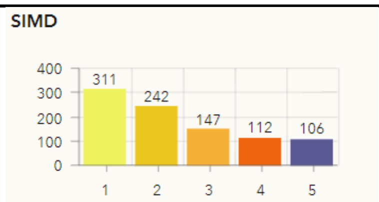
The table above presents the regression result. Two models were created in order to compare and choose the most robust model. Model 2 was chosen because it provides a model with a higher adjusted R^2 (0.80), the percentage of the dependent variable's variance explained collectively by the independent variables that imply the relationship's strength. The regression indicates that the active travel transport mode is related to accident rates, but with a small impact. However, regression results found that other factors have a more significant impact on the cause of accidents. First, the urban-rural class define the accident frequency, with the accessible rural area as the reference. In the regression model, we found that schools that are located in urban areas have a positive relation to the accident rate while the small town and rural area is the opposite. It means that the more urbanize schools' locations are, the more accident possible to happen. Second, the size of schools that are represented by the number of pupil rolls indicates a positive relationship to accident rates. This finding tells us that the more pupils a school has the more the accident rate could happen in its surrounding. Last, other variables such as the SIMD level and Intervention engagement are negatively related to the accident rate. From this finding, we could assume that the more developed the area will help to prevent road accidents around the schools. The intervention that has been implemented is also could be considered successful because the regression explains that the intervention engagement will reduce road accidents.

5.3 Results from Dashboard (query examples and insights from Dashboard)

There are multiple questions that can be answered with the dashboard, by filtering different variables and leaving others' contents. This provides the ability to gain instant insights. As examples of some questions that can be answered, this paper takes the two analysis questions explained before to compare if the quantitative analysis findings can also be visualized in the dashboard.

Do schools with accidents nearby have higher walking, scooting and cycling figures in HUSS? The findings were that active travel was positively associated with the number of casualties when controlling for other variables such as SIMD, intervention engagement, pupil roll, etc. To see if these findings were reflected in the dashboard, the casualty filter was used to look at how the other variables behaved with different rates of casualties. The following table illustrates the process by comparing two groups, one with the lowest half and another with the highest half:

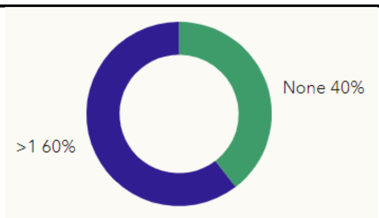
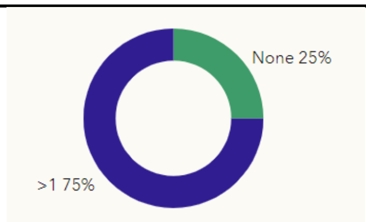
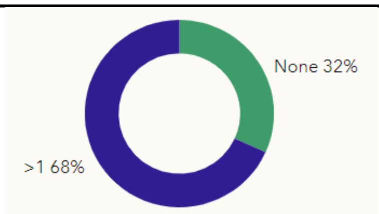
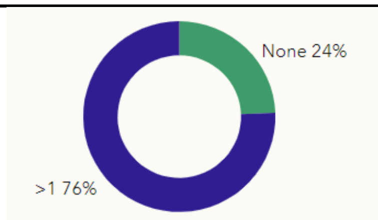
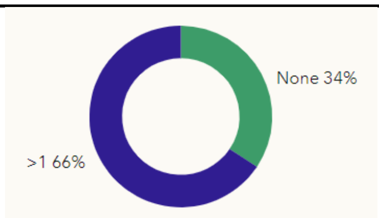
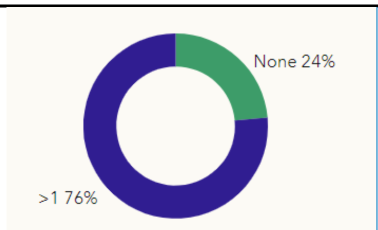
Table 7. Dashboard Findings of Accident Rate and Transport Modes

	Casualties: 1 to 5 Schools: 1085	**Casualties: 6 to 82 (max)** Schools: 918
Active Travel		
SIMD		

As seen in the table, active travel choice seems to be correlated with the number accidents, confirming the analysis results. Also, the SIMD was found to be negatively correlated with the number of accidents, which is depicted in the bar chart: among the schools with high casualty rates, there are more than double the schools in the first quintile than in the fifth.

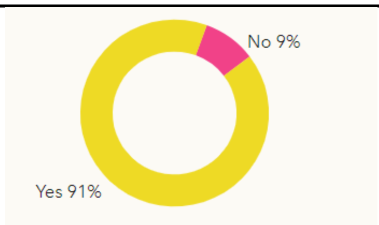
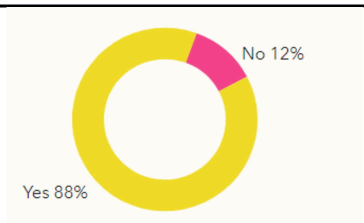
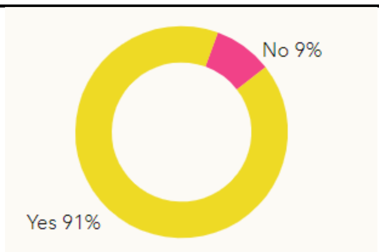
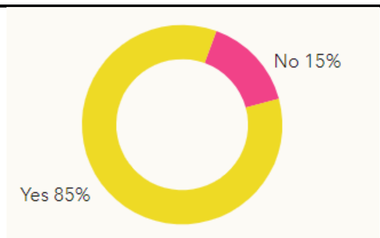
Schools that don't do HUSS and don't engage in interventions, are there any common characteristics? According to the analysis, characteristics that affect the intervention engagement are the urban-rural classification (positively), active travel (positively) and the size of the schools in terms of pupil roll (positively). The dashboard shows similar results:

Table 8. Dashboard Findings of HUSS Primary Schools Common Characteristics

	Filter 1	**Filter 2**
Rural Urban Classification Filter 1: Only Remote areas Filter 2: Non remote areas		
Pupil Roll Filter 1: Lowest half (<176 pupils) Filter: Upper half (>176)		
Active Travel Filter 1: Less than average <42% Filter 2: More than average (>42%)		

When looking at the increment of the influence variables, the proportion of schools that have engaged increased, proving the insights from the regression models. On the other hand, the characteristics that seem to be common in schools that have taken part in HUSS are the low number of casualties and the rural-urban classification. These are checked again in the dashboard and found the following:

Table 9. Dashboard Findings of HUSS Primary Schools Common Characteristic

	Filter 1	Filter 2
Urban Rural Class Filter 1: Only Remote areas Filter 2: Non-remote areas		
Casualties Filter 1: less than the average (< 11) Filter 2: More than the average (>=11)		

The charts are again aligned with the analysis findings, depicting the reduction in HUSS participation when looking at certain areas or a number of casualties respectively.

6. Conclusion

The project consisted first of condensing data from many different sources (i.e. non-profit organizations and different local, and national government departments) and from different fields (i.e. school poverty funding, SIMD, road safety, the spatial location of schools, HUSS Survey and records from active travel conducted at schools) to build a base dataset that allowed further data manipulation to create new variables on school road safety, the average modal choice for the past years or engagement in relevant active travel activities. Harnessing this data, two analyses were conducted through quantitative methods to provide some insight to Sustrans regarding school behaviour towards transport activities and road safety impact on mode choice. First, it was attempted to find common characteristics of schools that haven't shown participation in either the national transport survey or any active travel intervention. From the data available it was found that schools that are in remote areas tend to not participate in answering the HUS survey, surprisingly. Those same areas have participated more in active travel interventions which might be because they were intentionally targeted. Also, active travel share was associated positively with the interventions which might prove a positive impact of those efforts. In a second analysis looked at how the number of casualties that have occurred around schools is related to modal share or other characteristics, it was shown that active travel (i.e. walking, cycling or scooting to school) has a positive relation to the number of accidents when controlled for SIMD, urban-rural classification, and other factors. Moreover, it proved that the deprivation of an area increases the likelihood of casualties.

The main objective of the project was to produce a dashboard that would allow Sustrans to identify easily which school to target to have more positive outcomes when implementing an initiative or conducting an active travel intervention. Instead of showing only schools that correspond to specific thresholds Sustrans had set, it was proposed a flexible dashboard where they can decide on different criteria. The dashboard was built with the ArcGIS Online platform and it works by querying information from the feeding dataset using filter features corresponding to each relevant variable and then displaying that information about the schools in two components: a map, which contained schools' locations along with their detailed data, as well as all child casualties in Scotland. The second component was a set of charts summarizing data on participation, SIMD and active travel, which also provided the ability to instantly look at the influence of different variables on others and therefore get some insights. In the project, it was demonstrated that it was possible to instantly visualize in the dashboard the conclusion obtains from the analyses. Moreover, Sustrans will be able to export a table in CSV format of the queried schools to further work on them.

Future improvements can be made both in the dashboard and the analyses by gathering more data about the built environment around schools (i.e. proximity to cycling lanes, slope, residential or mixed areas, etc.), or any other relevant to understanding the interaction each school has with its surrounding. This way more detailed selection of schools can be achieved when deciding exactly what intervention is convenient for a school or to address which will have even more impact.

References

Bolboaca, S. D., & Jäntschi, L. (2006). Pearson versus Spearman, Kendall's tau correlation analysis on structure-activity relationships of biologic active compounds. *Leonardo Journal of Sciences*, *5*(9), 179-200.

Ewing, R. S. W. a. G. W., 2004. School Location and Student Travel Analysis of Factors Affecting Mode Choice. *Transportation Research Board,* 1895(1), pp. 55-63.

Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

Miller, M. (2021). Spatial Analytics Dashboards: Assisting Researchers to Select the Appropriate Tool. Bulletin - Association of Canadian Map Libraries and Archives (ACMLA), (167), pp.43–47.

S. Boniface, T. S. S. J. W. J. S. M., 2015. Health implications of transport: Evidence of effects of transport on social interactions. *Journal of Transport & Health,* 2(3), pp. 441-446.

Sidharthan, R. e., 2011. Model for Children's School Travel Mode Choice: Accounting for Effects of Spatial and Social Interaction. *Transportation Research Record,* 2213(1), pp. 78-86.

Sustrans. (2022). *About us - Sustrans.org.uk*. Retrieved 02 18, 2022, from https://www.sustrans.org.uk/about-us/