# CSCM27 Group Analytics and Visualisation of Information Project: Road Safety Accident Data for Wales

**Amal Abdulkader**
944093@swansea.ac.uk

**Anna Carter**
1915415@swansea.ac.uk

**Andrew Gray**
445348@swansea.ac.uk

**Ben Lloyd-Roberts**
880936@swansea.ac.uk

**Connor Rees**
872245@swansea.ac.uk

**Thomas Tasioulis**
998273@swansea.ac.uk

## ABSTRACT

With the increase in quality and scale of open source data repositories, understanding, interpreting and conveying such data continues to become an increasingly demanding task. This paper engages this challenge in using the UK Government's open source 2018 Road Safety Accident Dataset. In this context, the implementation of effective data visualisations becomes an integral factor of the project. Through employing effective visual analytical principles and methods, a subset of this large data set is effectively visualised. The said subset refers to the UK wide data being limited specifically to Wales. This dataset is, thereby, visualised into interactive coordinated multiple views. In such a way that is effective for a broad user base, from Governmental experts to stake holders to the wider public. As a result, filling a gap in the literature by using multi-view systems to visualise road traffic data. In doing so uncovering the risk associated with where you drive, when (per hour, day, month and year), road types, weather conditions per casualty type within Wales. Consequently, addressing the potential justifications of said findings, addressing bias and proposing the safest factors to drive in.

## INTRODUCTION

Personal injuries as a result of road traffic accidents are a significant public health concern. In seeking to preserve public safety in addressing this health concern. There is a need to demonstrate clarity and transparency through exploring and visualising large data on this topic. In working with a large dataset containing thousands of instances and tens of attributes, there is significant potential for visualising complex data in a way that is palatable to a broad spectrum of users. Ranging from experts in Government to various stake holders and the wider public. Thereby, the dataset that will be evaluated throughout this report is the 2018 Road Safety Accident Data. A UK Government, open source date repository which has three UK-wide road safety subsets on vehicles, casualties and accidents. The data has been collected over a significant amount of time, from 1979 to 2018. However, the data shape in this project has been restricted to consist of 4212 instances and 32 attributes limiting the data to the Wales region. For the purpose of this report 14 attributes have been used from the original dataset and 7 have been manually inputted for visualisation and analysis purposes. The attributes range from the severity of injury, vehicle type, road type and weather conditions. The added attributes include the principal components for Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). An attribute to enable the merging of Date and Time was also added.

However, the data set only includes accidents reported to the police or recorded using the STATS19 accident reporting survey form. This form of data collection could lead to potential bias as all of the accidents that occurred may not have been recorded [18, 17]. This report is going to focus on the sub-set of data which provides all of the accidents data for 2018. Due to the large volume of data available this report is going to focus only on the data collected for Wales. This is due to a number of factors, namely, due to hardware limitations, with the full dataset being too large to render the respective visualisations. And secondly in response to the need to improve road safety within Wales because of the increasing volume of accidents over the past few years. The data set can be found via the following link: Road Safety Dataset.

### Motives and Importance

The primary motive for this report is to bridge the gap between the general public, Welsh Government and national organisations who are attempting to reduce the volume of accidents occurring with the computers that are storing the data and detecting themes within this large data set. Visualisations can be used to bridge this gap because they enable users to view a vision that can represent the world around them on one screen [33]. They are also useful because they can produce highlighted relationships throughout the data which can be used as an incredible support for data analysis [31]. These creative displays aim to enable the users to evaluate the dataset and be able to suggest hypotheses about the possible correlations and relationships without the forcing of ideas and having to disregard outliers [21, 40]. Finally, visualisations are the best way in which to analyse this dataset as it prevents the need for reading through large volumes of data in a table setting and trying to link different attributes by eye [25]. Each of these factors entice our intended audience and bridge the gap between human and computational capabilities in interpreting complex and large data.

The number of people seriously injured by crashes within Wales increased by 7% from 2017 to 2018, resulting in 103 people being killed [22]. The Government have therefore set a target to reduce the volume of serious injuries caused due to crashes every year by 40% by 2020 [22]. However, with five people dying every day due to fatal crash injuries across the UK, this could be an unrealistic prospect [10]. Brake [10] is a charity based in Wales that aims to raise awareness for road safety. They found that 44% of the total deaths that occurred in 2017 across Wales occurred due to fatal accidents [10]. Furthermore, they discovered that an individual involved in a motorbike accident is 55 times more likely to be involved in a fatal accident than any other vehicle type [10]. However, things may not be as bleak as they seem, due to a trend in accidents over the last five years reveals that the number of fatal accidents are beginning to plateau with little variation from year to year. Although this may appear to be a significant positive, it can be seen to be a disappointing result due to the substantial decrease of almost 50% of crashes from 1979 to 2008 that were documented. [17]. The charity Children in Wales [26] found that over $145,000$ children and young people have to attend an emergency department due to an unintentional injury. Therefore, they are trying to find the main causes of these injuries to try and reduce the severe effects on children's lives. Another motivation for this project therefore is to aid the charities and governments to reduce the number of casualties throughout Wales by providing insights into the most common instances at which accidents and fatal casualties occur. This will also help to decrease the lost quality of life, estimated to be 2.5% of the GNP [15].

### Aims, Goals and User Scenarios

There are two main aims that this report aims to address. The first is to try to follow Schneiderman's [38] visualization mantra which consists of overview, zoom, filter, details-on-demand, relate, history and extracts. The second aim of this report is to provide an overview of the accidents and casualties that have occurred across Wales within 2018. This is to enable us to interact with the visualisations that have been produced and evaluate the themes and trends that emerge as a result. This overview will contain the ability to filter certain instances and retrieve details on demand to enable analysis on the dataset. The primary goal of the report is to find a pattern between the three injury severity types from the accidents and the several causalities. This will then enable us to provide feedback that Governments and charities will be able to base their safety adaptations on. This will include the worst areas, weather conditions and speeds at which fatal casualties occur. Finally, there are four user scenarios that will be covered within the report and are shown below:

1. What is the worst time of year for crashes? (Month/Day/Hour)?
2. What is the worst type of road for crashes?
3. What are the worst weather conditions to drive in?
4. When is the safest time to drive?

## PREVIOUS RESEARCH

There are five key themes that can be derived from surveying the relevant research on this topic. These consist of over representation of clustering methodologies, single car accidents, young driver research, the juxtaposition between road safety research being based on traffic and finally the importance of using maps for spatial data.

The first is the use of clustering to evaluate road safety datasets which has been used to find that Support Vector Clustering methods use both spatial and temporal cluster [32]. [3] uses this process to evaluate the accident hot-spots within an estimation map and believes it is an effective strategy for the reduction of high density accident areas. This approach is similarly used by [36] who uses K-means to cluster different types of traffic conflicts from video-graphical data. In contrast to this [37] used clustering analysis to evaluate the similarity in accident prevention practices between the different countries across the globe to provide recommendations for the countries that are seen to be lacking in policy on how they can improve. To try and contradict this trend [32] used spatio-temporal interaction to see the effect of vehicle crashes across the UK which provided a much needed insight into the variance in crash causes. However, this only provided a small insight of the dataset that was evaluated and therefore this report hopes to increase this overview more broadly for Wales.

Secondly, a vast amount of research has been reviewed in is with single car accidents. For example, Erdogan [16] evaluates the correlation between the spatial distribution of the dataset and the death rate of the casualties in Turkey obtained using single car accidents. They found that the largest death rates involve the roads connecting Turkey to Istanbul, Ankara and Antalya. This approach is also taken by Huang et al. [24] who evaluated how a single car and pedestrian could be used to produce a sensor system that could be implemented on cars to detect pedestrians on the drivers behalf in Sweden. However, these reports did not recognise or evaluate the multi-car crash accidents which could have led to bias within their work. In contrast to this [27] found that single and multi-car crash rates often have very different attributes contributing to the severity of casualties including traffic intensity and shoulder width of the road. [28] also found that traffic intensity and light conditions effected the accident severity very differently for single and mutli-car crashes. Therefore, we aim to use both single and multi-car crash data to try and prevent bias from the dataset and from narrowing our results to only particular instances.

Thirdly, a vast amount of research has been completed into the affect of age on both the cause of casualties and the casualties that occur with results mainly relating to young drivers. This may not be deemed a surprise as the World Health Organisation (WHO) [30] reported that the current leading cause of death for children and young adults between the ages of $5-29$ is due to road traffic injuries compared to the eight for overall age groups. The main reason for these incidents is due to risk taking and mobile phones including the issues with excessive speed, loss of control and the failure to detect other vehicles being a common correlation across many reports [34, 9, 12].

However, work has also been completed on the increasing elderly population. For example [6], completed user testing within the USA with older drivers to evaluate the damage of memory loss and slower reactions on crash probability using a simulator and found that there was a large correlation. This area of research is highly interesting however the ages of the individuals within the dataset have been removed to ensure anonymity. Therefore, this is something that will not be addressed within this report.

The fourth key finding drawn from surveying the literature is two broad themes which stem from road safety research. The first stems from the research previously referred to in this chapter which is road safety research referring to accidents. A large collective of the remainder of the data addresses road safety from a traffic reduction perspective [35, 42]. Largely referring to the accurate clustering of spatio-temporal data as opposed to exploring multidimensional data with geo-location. This is just one of a multitude of attributes that can be explored.

Finally, the importance of visualising the spatial data using a geographical map was made clear [4]. The idea of a 'comap' is to include both a background outline of a geographical map and the spatial data points onto the map to enable the evaluation of different instances in different locations [11]. This type of evaluation can also be used for spatio-temporal analysis such as evaluating fire incidents across a certain location [13]. This method is beginning to be widely implemented throughout research in the use of road safety visualisations. It enables the classification for the volume of accidents occurring in certain locations and these locations can be evaluated further using zooming into particular counties or countries within the map [23, 19]. This visualisation would also be helpful in evaluating commuting distance and the number of casualties. However, currently only regression models have been used [20, 29]. We hope to implement a comap within our report including all of the spatial data across Wales and relate this map to other graphical visualisations which is not widely covered across the literature.

There are several other approaches that have been completed within the research that have attempted to fill different gaps within the research to a particular extent. [34] tried to alleviate the bias within the dataset by asking both police officers and members of the public to analyse the main causes of road accidents reported in the accident records. However, it has been found that several declassification's can be made within police data when reported from a crash [2].

However, another approach that has been considered is to evaluate the time of day in which the biggest number of accidents occurred. For a group of 3000 accidents across mid-Britain it was found that accidents were at their highest during the evening when there was less available light [12]. This hypothesis was further analysed across the year and it was found that this value increased further in the summer months with causes being related to sleepiness and alcohol intake [1]. However, there has also been a relation found between a large volume of accidents occurring during the afternoon and this was found to be related to the behaviour of the driver [43]. A variety of different types of data we explored to evaluate this hypotheses including video data where cameras and sensors were used to evaluate the behaviour of the driver when accidents occurred. It was found that 90% of crashes were caused due to fatigue, distraction or error of the driver [14]. One of the few reports that produce user testing into collision avoidance produce a prototype that implements collision avoidance technology. However, it was found to be ineffective and many users continued with their previous driving habits [39]. Although the research is lacking in relationships between the time of day and weather conditions, road types, locations across the UK we hope to reduce this gap.

Several areas that we hope to address which we have not found to be widely researched include visualisations that provide clear areas in which preventative measures or recommendations that can be implemented to contribute to reducing the number of casualties. Evidencing a requirement for research papers to step beyond the realms of academia and make real-world recommendation based on their respective findings. In conjunction, what has been highlighted is the apparent lack of visualisations on road safety accidents to adopt Shneiderman's visualisation mantra. With the work failing to adopt the overview first to details on demand. Existing literature is restricted to overview with few examples including zoom and filter. As a result this paper aims to counteract these limitations by using multi-view systems that are interactive with each other.

## IMPLEMENTATIONS AND RESULTS

### Final Implementation

Figure 1 shows the final representation of the dataset using coordinated multiple views. A multi-view system was used as it is one of the most successful ways visualising high dimensional data enabling the user to view as many of the attributes at once [7]. The ability to evaluate many different views within one screen has been proven to be highly effective as it supports short term memory by filtering and highlighting chunks of information within the different views [8].

Figure 1*a* shows the heat map with hours of the day plotted against the months of the year. The red saturation gradient encodes the frequency of road accidents over 2018, lighter sections denoting the safest times of travel. Red saturation was selected due to the human brain's pre-attentive process which subconsciously identifies red colours [44]. Considering our solution aims to inform its users of road risks, drawing attention to the time frame with the greatest potential for accidents to occur was essential. The heat map can be highlighted with a brush cursor to filter accident information by month or by hour. This enables the user to evaluate temporal patterns by deriving information from the linked views. A tooltip function was also included so users could query individual cells within the heatmap to view the total number of instances are to be view within that time frame. By incorporating this functionality in to the heatmap, both colour and discrete values inform the user of accident frequency for a given time frame.

Figure 1*b* shows the casualties that occur by each week day with colour relating to the severity of the casualty. We included this view to accompany the heatmap overview of the year. This

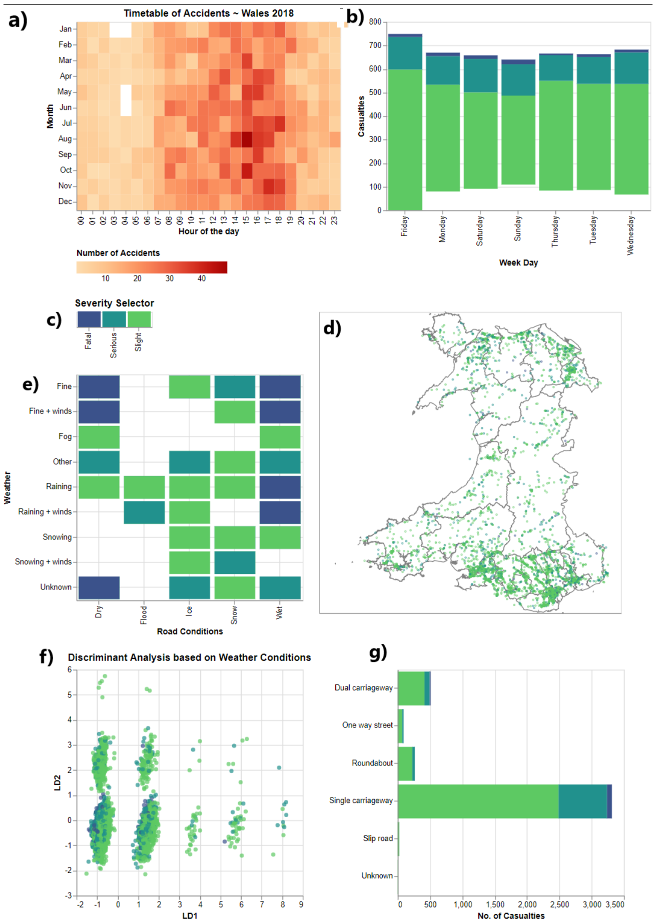

**Figure 1. Final implementation of interactive, linked visualisation suite. Coordinated multiple views were employed to better represent the high dimensional, multivariate dataset.**

way users are capable of identifying weekly and daily trends according to season or time of day.

Figure 1*c* shows the severity selector view. This selector can be used to filter the different attribute information by accident severity. As a different severity is chosen, correlations can be evaluated by simultaneously updating other coordinated views. Accident severity is graded one of fatal (Blue), serious (Teal) or slight (Green). Interaction supports single or multi-select to filter by individual or multiple severity classes.

Figure 1*d* shows a spatial data map also known as a conditional map (comap) which contains all of the points plotted on top of a geographical representation for Wales. Map interaction allows users to select regional information by highlighting areas of the country with a brush cursor. All linked views are concurrently updated with the attribute data for the highlighted region. The extent of linked filtering is further discussed after Figure 2.

Continuing down our visualisation views, Figure 1*e* visualises a grid of road surface and weather conditions, also uniformly coloured according to the accident severity scheme. These tiles are filtered according to the heatmap selector, map brush and by selecting the tiles themselves. Each grid cell represents a sub-category of all instances that occurred during the respective driving conditions. Users may 'peek' at the composition of these categories by using a tooltip to interact with each cell. Once selected all other linked views will filter according to the selected driving conditions.

Figure 1*f* shows the Linear Discriminant Analysis (LDA) for the weather conditions attribute, this graph can be highlighted to view how the different types of weather conditions effect the rest of the attributed views. Finally, Figure 1*g* shows the correlations between the number of casualties and the road type again with colour represented as the severity of the casualty. Once again users are able to hover their cursor above each bar to view the total number of casualties within that class of road type.

From an overview of the software some key insights were found. It was clear to see that the majority of the accidents happened within South Wales. Especially in the key cities of Cardiff, Swansea and Newport. However, for the North of Wales accidents did not tend to occur around main towns. This could be due to the population of North Wales consisting of many small villages situated relatively far apart.

Another clear finding from the visualization was that most of the casualties occurred on single carriageways. A total of 3,313 casualties were recorded to have occurred on single carriageways, with the second highest volume of casualties being 509 for dual carriageways. The smallest number of casualties were recorded on slip roads. However, all of these casualties were fatal. The large volume of casualties that occur on single carriageways are clearly an area of great need for creating new safety regulations. This could be due to many single carriageways being narrow, with fast speed limits and often very little lighting.

Another insight that can be taken from the initial visualisation is that the most common day of the week for an accident to occur is on a Friday, where a total of 749 casualties were recorded across the year, almost 20% of the total accidents number. It appears from the visualisation that Sunday is the safest day to drive with a value of 378. However, many of these accidents resulted in fatal casualties.

Figure 2 describes the binding and filtering relationships between each interactive component. It is the intention of Figure 2 to better illustrate the overall filtering capabilities of our proposed solution. Once the implementation of our solution had taken place, a rudimentary sketch of the above graph was generated to identify the in-degree and out-degree of each view. The in-degree of a particular view, or node, was calculated by the number of interactive views capable of filtering the data shown within a single view. The out-degree of any given view was determined by the number of external views that the single view was capable of filtering. For example, we observe both 'Severity Selector' and 'Heatmap' views sit at the top of the interactive hierarchy, each with 4 views they are able to filter. There are also no other views that are capable of controlling the information that they display. One-way filtering relationships are represented is by unidirectional arrows

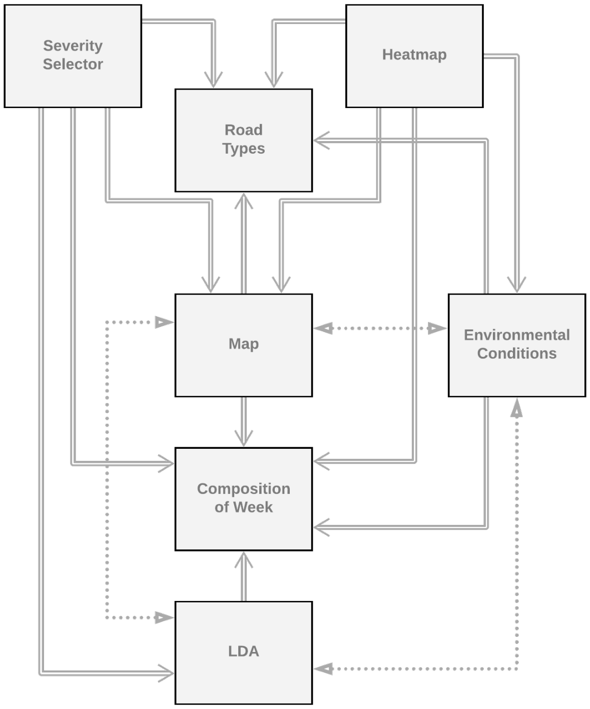

**Figure 2. Graph representation of binding relationships between the interactive views in our solution. Dotted lines represent binding relationships with two way filtering and interaction. Solid lines denote a one-way filter relationship.**

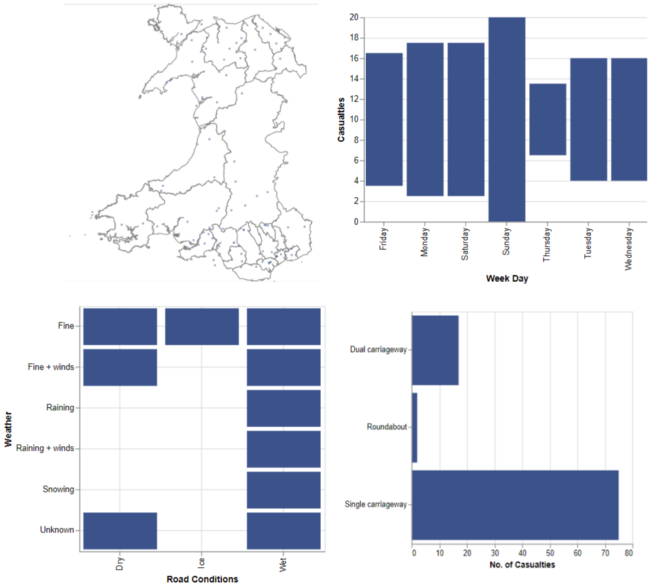

**Figure 3. Showing the evaluation of the selection of Fatal accident severity on the remaining multi-view graphs.**

which begin at the 'controlling' view and end on the filtered view. The views are capable of applying filters to each other and are represented by double ended arrows joined by a dotted line.

## Casualty Patterns

In this section we aim to evaluate the filtering of data using the severity of casualties to identify any underlying data trends. For each figure, a different severity was selected to evaluate how the rest of the visualisations filtered in relation. This was pursued to identify causality patterns between the severity of crashes and other attributes. The attributes updated according to the severity selector were the week patterns, the location of the incident, the driving conditions and road type by casualty count. All of these factors will be measured against the three incident severity's: fatal, serious and slight.

The first attribute to be discussed is the week day which reveals numerous findings worthy of note. As can be derived from the bottom right chart in Figure 3 which compares the week day to the casualty count, partitioned by severity, two broad findings can be derived. Firstly, the majority of incidents take place over the working week days (Monday to Friday) and secondly, the most fatal incidents take place on the weekend. Firstly, the majority of incidents take place over the working week days is a statistic heavily weighted by the slight severity instances. Slight accidents contribute to more than the fatal and serious categories combined, and this is the case for every day of the week. However, the number of casualties was the highest in weekdays, with Fridays tied to the highest frequency with 598 casualties. Comparatively, fewer incidents occur on Sundays with 378 recorded case during 2018. The remaining days of the week average 450 casualties each. In reference to the statement that fatal casualties were more prominent on the weekend; Sunday in particular had the most fatal accidents totalling 20 fatalities, with Saturday evidencing a similar result at 15 with over 80% of the accidents occurring on a single carriageway. This trend becomes more evident with this gradual increase from the week to the weekend with Friday being the third most common day for fatal casualties with 13 instances.

The availability of geolocation data is integral to the implementation and the coordinated multiple-view structure. The map view which illustrates Wales with data points plotted over the top provides clear representations for observations and findings to be drawn. The map views shown in Figures 3, 4 and 5 display data that can be separated into distinct clusters. In reference to figure 3 in particular, plotting all of the instances on the map view illustrates areas with densely populated instances. What becomes immediately clear in the dense instance clustering in South Wales and North Wales is that city centers (where these dense clusters preside) are more prone to accidents than other environments [41].

The remainder of the incidents are sparsely distributed across the map which is a finding that may be correlated to the smaller population of areas such as mid Wales. This data can therefore be segregated into mainland Wales and coastal Wales. However, comparing the North and South regions of Wales proves to be more fruitful. For example, in separating the data in this way, it could be construed that incidents are significantly less prominent in North Wales in comparison to South Wales. With North Wales containing a total of 37 fatal accidents, 368

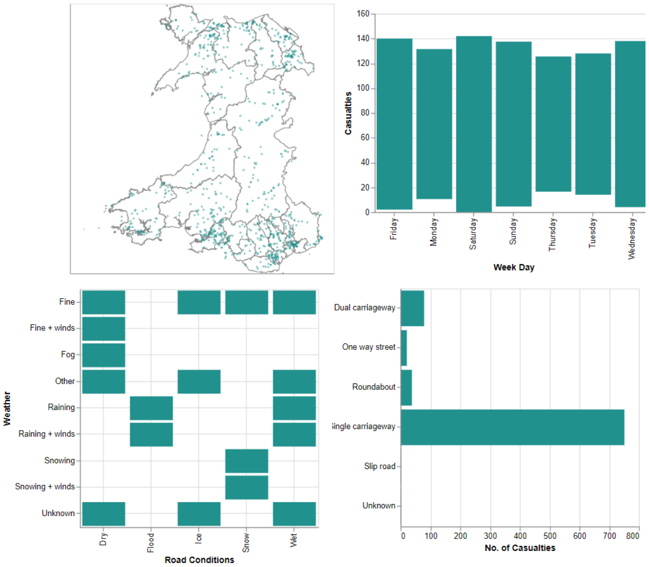

**Figure 4. Showing the evaluating of the selection of Severe accident severity on the remaining multi-view graphs.**

serious accidents and 918 slight accidents. In contrast, it was concluded that South Wales has a total of 58 fatal accidents, 525 serious accidents and 2321 slight accident severity instances. In exploring this data, it was noted that there were 1080 fewer accidents on single carriageways in North Wales than South Wales. However, the average fatal incident severity level is significantly higher for North Wales. As a result, in exploring the differences in incidents between locations, a potential area for exploration in the data was the finding between incidents and road types. Thus, this is consequently explored in the following subsection.

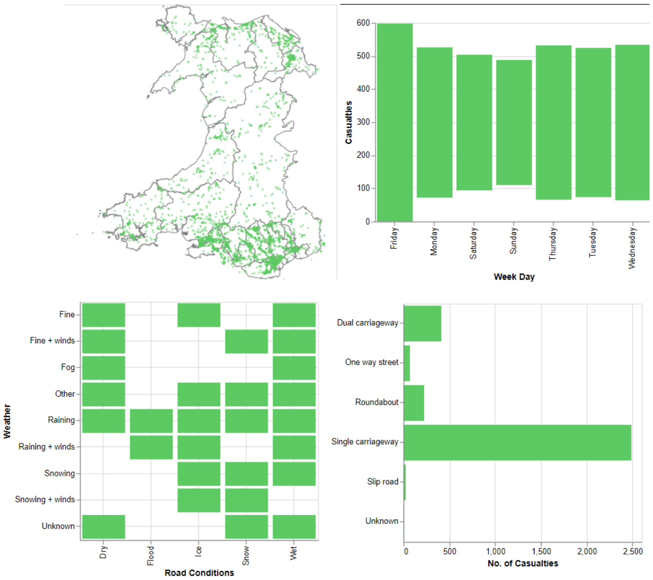

**Figure 5. Showing the evaluation of the selection of Slight accident severity on the remaining multi-view graphs.**

We can observe from all three severity classes that their attribute composition does not alter much other than casualty count. The road and surface condition view in Figure 3 does provide some insight into a potential seasonal influence on the accident severity. We note that there are far more cases of fatal accidents occurring on wet road surfaces than any other condition. As severity begins to decrease, the variety of driving conditions in which road accidents may take place also increases. This trend matched expectations considering the likelihood of slight accidents is significantly higher than the occurrence of fatal collisions. This may also inform us of a potential randomness to the occurrence of slight or serious incidents. Attributes such as weather and road surface conditions appear to have less influence over these severity classes, particularly in comparison to fatal cases.

### Dangerous times to drive

Throughout this section the heat map and week view will be used to highlight different times of year including hours, days and months within the dataset. The results on corresponding graphs will then be evaluated. From Figure 6*a* numerous findings pertaining to accident frequency based on the hour of day framed against the twelve months of the year can be seen. The visual overview clarifies several key themes about particularly safe and unsafe times to drive. The increased saturation of the red seen in what may be summarised as a dark red vertical cluster clearly communicated that between the hours of 3pm and 5pm is consistently the most accident prone time frame for road traffic incidents throughout the entirety of the year. This may be a byproduct of children leaving school at the beginning of this time window. Additionally, the other end of the time window corresponds with the concept of rush hour, whereby, the average employee finishes work and begins their commute home.

Although this theme remains consistent throughout the year, the occurrence of incidents in these windows and in the rest of the day decreases from November through to March. This may be due to better weather conditions in the spring and Summer months giving drivers a false sense of security. It may also be due to the population increase during these seasons as a result of tourism. Conversely, the heat map visualisation also identifies that from late in the evening to early in the morning (from 9pm to 6am) is the time window that sees the least occurrence of vehicle accidents. The most likely justification for this is the significant reduction of congestion and thereby reduced opportunity for human error which is so readily attributed to vehicle accidents.

Figure 6*a* shows that the largest volume of crashes across the year were found to be in August. During August, the system shows that Friday is the highest day for casualties. This amount reflects the overall trend for Wales. This is because a total of 9 fatal casualties occurred in August. Figure 6*b* shows that for the month of August 3 fatal casualties occurred on a Sunday and 2 on a Monday. The comap showed that North Wales had a total of 6 of these fatal casualties. Therefore, it could be concluded that North Wales has a tendency to have more fatal accidents during the tourism period than the remainder of Wales.

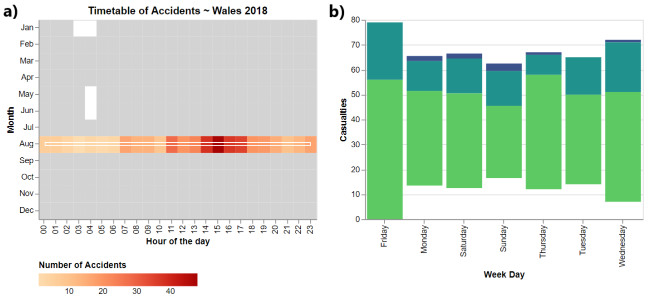

**Figure 6. Showing the evaluation of the increased casualties occurring in August compared to the remainder of the year.**

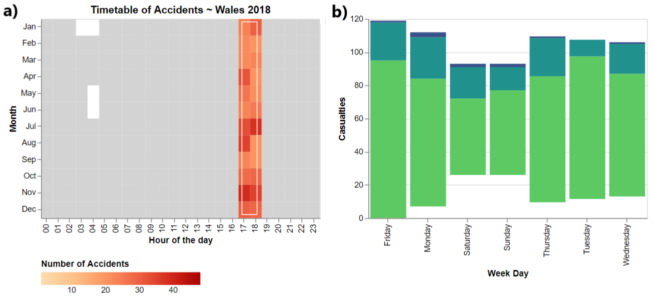

**Figure 7. Showing the evaluation of the increased casualties occurring at rush hour compared to the remainder of the day.**

Figure 7*a* shows that a large volume of casualties occur at rush hour across the year. This is not a surprise as there is often large volumes of congestion due to commuters. Figure 7*b* shows that the number of casualties occurring between these times are significantly reduced on the weekends. This could be due to commuters often working a Monday to Friday job and therefore the volume of congestion at these times would not be as severe.

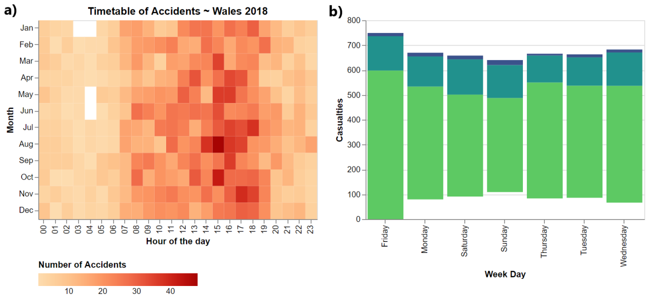

**Figure 8. Overview of temporal data across** 2018**.**

Figure 8 reveals the most accident intensive time frames throughout 2018. As expected this ranges across a 12 hour window from 7am to 7pm, which corresponds to common working hours and commuting times. We also observe the safest times of travel, most notably during unsociable hours, early in the morning and late at night. This is likely attributed to the reduced congestion of roads during these hours. It is also worth noting how this corresponds to a fairly even distribution of accidents over the week, with the exception of Friday. The composition of those days by accident severity is fairly uniform. This is further explored in Figure 10.

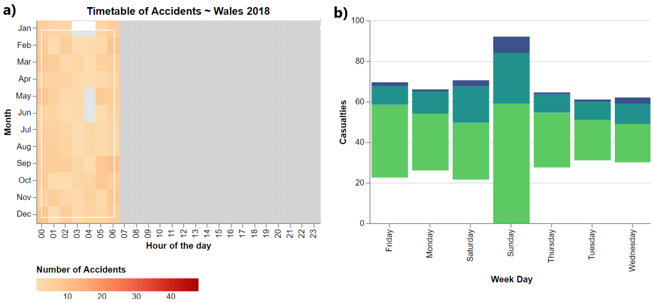

**Figure 9. Filtered heatmap according to safest time of travel, deemed by the time frame with the fewest road traffic accidents.**

Figure 9 demonstrates the effect of isolating the time of day with fewest incidents, $00:00$am to $06:00$am respectively. We observe a sharp increase in the number of accidents occurring on Sundays. Due to our ability to highlight select hours of a given day, we are able to identify potential reasons for the incident spike on a Sunday. While technically these incidents occur on a Sunday, this is likely a continuation from incidents occurring on Saturday evening.

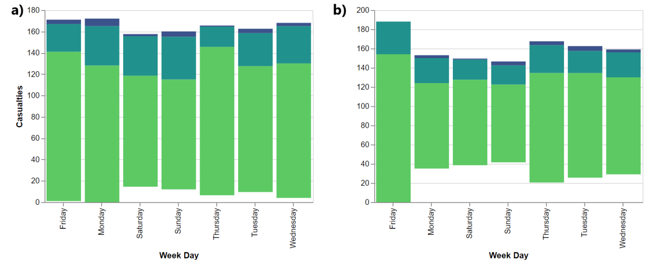

**Figure 10. Comparison of accident composition over the average week during winter and summer seasons. Figure a) representing the average weekly composition of traffic accidents according to the number of casualties and their severity.**

The distribution of incidents during the average week from June to August can be derived from Figure 10*a*. The corresponding Figure 10*b* exhibits the average distribution of incidents during a week from January to March. We see a pattern shift in terms of which days most accidents occur, namely weekends are safer times of travel during the winter. In stark contrast summer months appear to be comprised of a more uniform distribution of accidents across the week. During a prolonged holiday season, work schedules are less restricted to the typical $9-5$ work hours, meaning travel can take place at any point in the week. We also observe that regardless of the season, the composition of a given day by accident severity remains the same, they're just more likely to happen during warmer seasons with tourism being synonymous with travel.

## Bad weather conditions

Determining the safest and most dangerous driving conditions is a conclusion that is drawn in reference to the accident severity. As displayed at the bottom left of Figure 1, the road conditions and respective weather conditions are plotted against one another. However, due to the data being largely made up of slight accident severity, the visualisation is less effective in portraying the weather conditions for each of the accident severity categories i.e. fatal and serious accidents. However, the interactive element of the visualisation circumvents this limitation of the initial overview.

Firstly, in looking at the fatal incident severity instances possibly the most foreseeable and comprehensible theme immediately presents itself. In that, so long as the road condition is wet (reducing traction of the vehicle to the road and thereby reducing the user control), the weather condition becomes a less significant factor, as seen in Figure 3. Having established this interesting result, it is, however, worth clarifying that wet road conditions and raining weather conditions did accumulate more instances than the rest of the weather conditionson wet road surfaces. With wet road conditions and raining weather condition being the second largest combination and fine weather conditions being the third most prominent. Thus, reinforcing the finding that wet road conditions are inherently the most dangerous driving conditions. This risk only increases when raining weather conditions are met in conjunction. However, possibly more surprisingly, the combination of dry road conditions and fine weather conditions (what one may presume as being optimal driving conditions) results in the highest fatality instances. Accumulating over three times the amount of instances of wet road conditions and raining weather conditions (the second largest combination).

Secondly, is the serious severity instances, which immediately appears broader in including more road conditions, weather types and range of combinations. This is likely due the increased number of instances which as a result will increase the likelihood of increased road and weather conditions. Despite the increased instance variance, the data begins to show several patterns. Firstly, regardless of weather conditions wet road conditions appears to be a significant factor in road accidents. In addition, the top three most prominent combinations of road and weather conditions remains consistent. With dry road conditions and fine weather being the most accident prone conditions followed by wet road conditions and raining weather and then wet roads and fine weather.

Finally are the slight incident severity's which account for the majority of the dataset. Similarly to the observations made on the serious accident severity instances, slight injury instances encompassed an even broader range of road and weather condition combinations. This again, may largely be attributed to the significant increase in number of instances above other factors. Additionally, the three most prominent combinations of road and weather conditions is the same for slight injury instances.

Figure 11 visualises a set of Linear Discriminant Analysis (LDA) clusters which represent our 14 attributes according to accident severity, road surface conditions and weather conditions. Highlighting individual clusters reveals the driving conditions shared among datapoints within that specific clusters. In Figures 11*a* and 11*b* it can be observed that selecting different clusters along the *LDA*1 axis, reveals that road surface condition is a discriminating class. Additionally, shifting the focus to clusters along the *LDA*2 axis reveals that surface

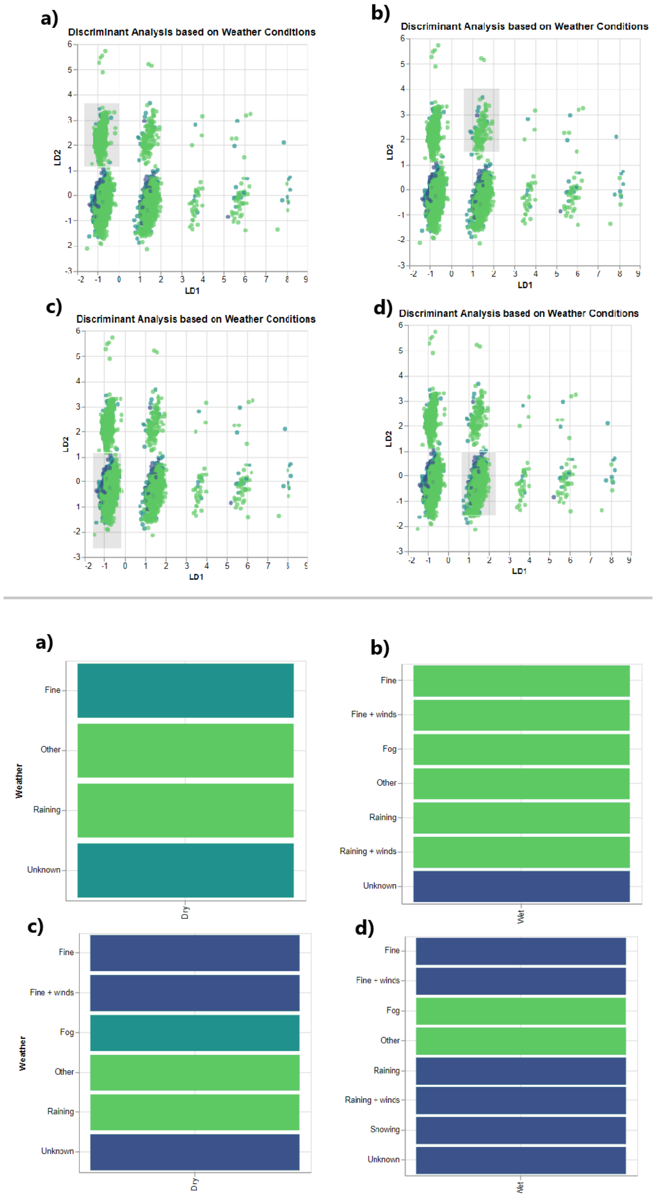

**Figure 11. Linear discriminant analysis was performed with respect to weather conditions to maximise variance between classes and minimise in-class spread.**

conditions remain the same yet, we are now isolating less severe accidents occurring across a range of weather conditions. Once this relationship is understood, users will be capable of quickly identifying the environmental conditions wherein most road accidents occur. As Figure 2 demonstrates, we also see that our LDA views are capable of being filtered by a number of other attributes. Clusters can be reduced down from a general overview to reveal the driving conditions for a specific accident severity, a particular time frame or by location. This procedure allows us to both view general trends in weather and road conditions as well as filter certain cases.

An immediate trend to notice in Figures 11$c$ and 11$d$, is the increase in the number of fatal accidents in wet driving conditions. This is to be expected, as the decrease in tyre friction along wet surfaces increases the difficult of stability and control of the vehicle, particularly when travelling at speed.

## Risky Roads

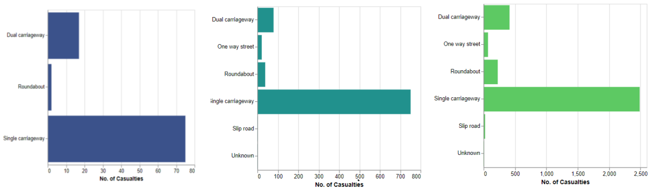

**Figure 12. Showing the evaluation of weather and road conditions on the three different types of casualty severity: fatal (left), severe (centre) and slight (right).**

Figure 12 represents the number of casualties occurring on a variety of road types, partitioned by accident severity. Immediately we observe that incidents, regardless of their severity, are most likely to occur on single carriageway roads. For those familiar with transport infrastructure in Wales, single carriageways are the most common road type across the country, this trend is to be expected. However, we also note that as accident severity increases there are fewer potential road types where these incidents can occur. Simply put, these figures suggest there is an increased likelihood of fatal cases on roundabouts, dual and single carriageways.

## Alternative Implementations

In this section, the alternative implementations that were explored to see how best to evaluate the dataset that was selected shall be discussed.

Visualisation is an important principle that is required for in-depth analysis of large quantities of data. In an attempt to support comprehension of the high dimensional data the intention was to generate a simplistic, broad overview of the entire dataset. As spatio-temporal information was readily available in the road safety dataset, the establishment of two essential views to include within the solution became imperative. With a significant number of instances recorded across the whole year, a visual component capable of both general overview and querying specific time frames was required. Figure **??**$a$ shows the first attempt at visualising annual patterns. However, in accordance with the Schneiderman visualisation mantra [38], the ability to filter information, zoom in on interesting features and acquiring details on demand became crucial. To this end, the stacked bar chart proved to be insufficient. This particular view was replaced with a heatmap representation of annual incident occurrences, partitioned by month and time of day. Additionally, the inclusion of a coordinated view to further visualise the composition of a given week within a selected period of time.

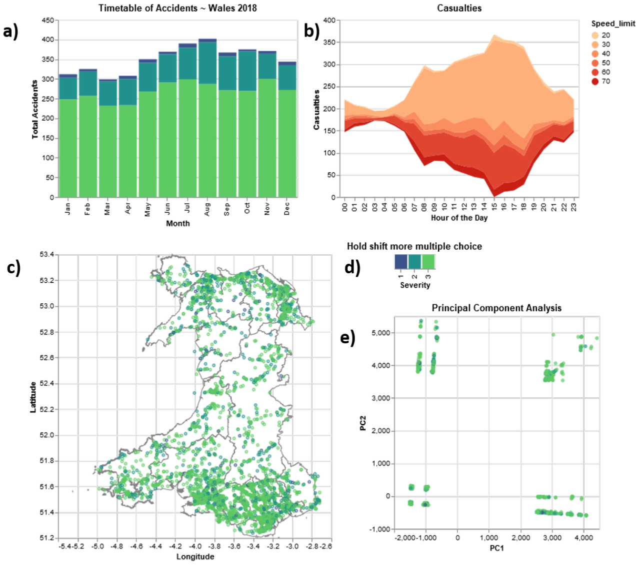

**Figure 13. Alternative implementation. Figure a) shows the monthly distribution of accidents against accident severity. Figure b) shows the distribution of casualties per hour with the colour saturation relating to the speed limit of the road. Figure c) shows the spatial data represented on top of a geographical representation of Wales. Figure d) shows the interactive selection of casualty severity. Figure e) shows the variance of the dataset using Principle Component Analysis.**

Figure 13$b$ explored the relationship between the speed of travel before the accident took place, the time of day and how this influenced casualties. While this provided an informative overview of the annual trends, the steam graph had limitations when many filters were applied and only a few instances remained for analysis. Figures 13$c$ and 13$d$ were agreed as fundamental components to the final visualisation suite and therefore no alterations were made. All road traffic data was accompanied by latitudinal and longitudinal positions for each instance, making a map representation vital for identifying dangerous regions. Similarly, an interactive view capable of filtering all information according to the accident severity seemed vital. For users of our solution aiming to avoid harm during travel, the ability to determine the potential risks of travel is paramount. This is useful as it enables the user to conduct more in depth investigation of the selected elements of the dataset. As such this component remained unchanged. Finally, Figure 13$e$ exhibits our initial attempt at using Principal Component Analysis (PCA) to reduce the dimensionality of the dataset and potentially reveal any underlying features. Unfortunately the resulting clusters visible in the alternative design did not provide ample information. Therefore, this requirement was ultimately not satisfied.

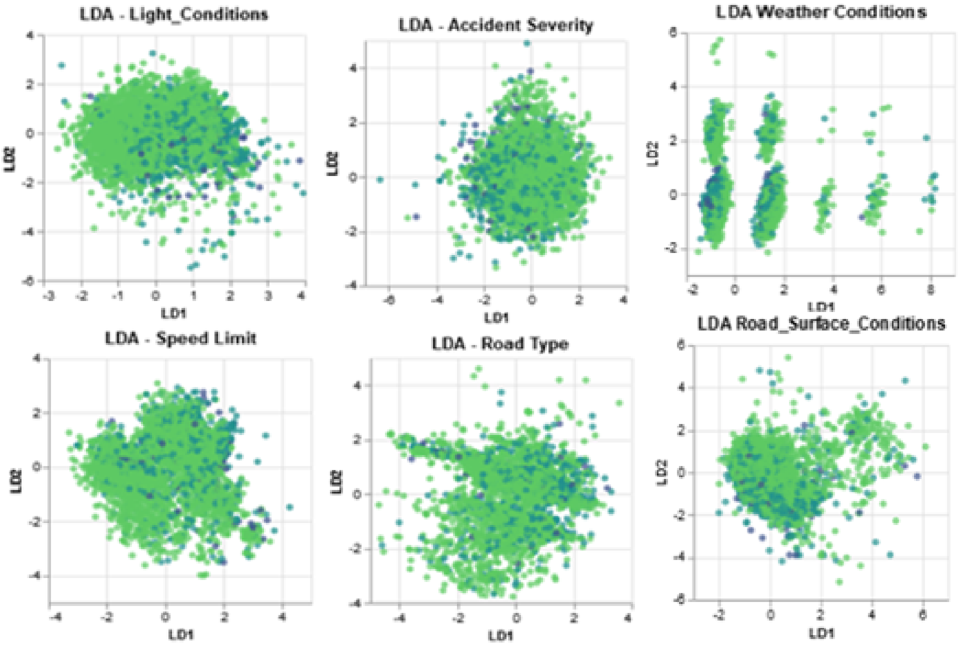

**Figure 14. Linear Discriminant Analysis of six different attributes, showing that the only helpful attribute from this selection is the Weather conditions.**

The second alternative method uses Linear Discriminant Analysis (LDA) to evaluate six different attributes: Light Conditions, Accident Severity, Weather Conditions, Speed Limit, Road Type and Road Surface Conditions. This alternative is shown in Figure 14 which shows that the only attribute that provides a helpful insight into the data set is the Weather Conditions attribute. LDA was evaluated because it is a dimensionality reduction method commonly used to maximise class separation [5]. As the weather conditions were the only insightful clustering of the data, this graph was added to the visualisation.

## CONCLUSIONS AND FURTHER WORK

To conclude, firstly, we believe that within this report we have addressed a gap within the literature surveyed. This was accomplished by providing an overview of the dataset using coordinated multiple views and view binding. We used both single and multi-car crash data to try and give a less bias overview of the data collected across the UK. The successful implementation of a map enabled the ability to view all of the spatial data collected across Wales in relation to our linked views.

Secondly, we have found answers to the four user scenarios that we set out within our aims and goals. It was found that the worst time of year for travelling within Wales is August due to the increased number of cars entering the country due it being tourist season. It was also found that the most dangerous time of day to drive was between 4pm and 6pm (rush hour). The second user scenario related to the worst type of road for crashes and it was found that single carriageways incurred the most accidents for all road types. Wet or snowing weather and road conditions were found to be the most dangerous conditions to be driving in with a high percentage of casualties. Finally, it was found that the safest time to drive in Wales is during October around 3pm. We believe that if the Welsh Government and local Charities take this information into account when analysing new ways to improve road safety they will be able to reach their goal of reducing road traffic accidents by 40% by 2020.

Future development of this work would lead to evaluation of road safety data across the entire United Kingdom. Currently, the UK government maintain several sister datasets to the one we have incorporated in to our solution. These datasets comprise of vehicle-related information tied to road accidents as well as demographic and casualty information for all recorded road accidents. The vast quantity of these additional datasets would allow for more revealing features produced by analytical techniques. Further operations such as spatial clustering on a map or feeding existing data into a learning model to predict future accident severity, location and time of occurrence. Many attributes were also made redundant by limiting our dataset to Wales. Population differences also reduced the quantity of information we were able to incorporate in to our solution. Information exclusive to densely populated urban areas was lost due to the rural nature of Wales. Population differences also reduced the quantity of information we were able to incorporate in to our solution, meaning trends observed in our solution may not necessarily translate to the UK as a whole.

## REFERENCES

[1] Kecklund G. Akerstedt, T. and L.G. Horte. 2001. Night Driving, Season, adn the Risk of Highway Accidents. *Sleep Research Society.* 24, 4 (2001), 401–406. DOI: http://dx.doi.org/10.1093/sleep/24.4.401

[2] Martin J.L. Chiron M. Amoros, E.M. and B. Laumon. 2007. Road crash casualties: characteristics of police injury severity misclassification. *Journal of Trauma-Injury Infection and Critical Care* 62, 2 (2007), 482–490. DOI: http://dx.doi.org/10.1097/01.ta.0000202546.49273.f9

[3] T.K. Anderson. 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. *Accident Analysis and Prevention* 41 (2009), 359–364. DOI:http://dx.doi.org/10.1016/j.jaap.2008.12.014

[4] T.C. Bailey and A.C. Gatrell. 1995. *Interactive Spatial data analysis (Volume. 413)*. Longman Scientific and Technical, Essex.

[5] S. Balakrishnama and A. Ganapathiraju. 1998. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing* 18 (1998), 1–8.

[6] Edwards J.D. Ross L.A. Ball, K. and G. McGwin Jr. 2010. Cognitive Training Decreases Motor Vehicle Collision Involvement of Older Drivers. *Journal of the American Geriatrics Society* 58, 11 (2010), 2107–2113. DOI: http://dx.doi.org/10.1111/j.1532-5415.2010.03138.x

[7] Callahan S.P. Crossno P.J. Freire J. Scheidegger C.E. Silva-C.T. Bavoil, L. and H.T. Vo. 2005. VisTrails: enabling interactive multiple-view visualizations. In *VIS 05. IEEE Visualization, 2005.* 135–142. DOI: http://dx.doi.org/10.1109/VISUAL.2005.1532788

[8] Roberts J.C. Boukhelifa, N. and P.J. Rodgers. 2003. A coordination model for exploratory multiview visualization. *Proceedings International Conference on Coordinated and Multiple Views in Exploratory Visualization* (2003), 76–85. DOI: http://dx.doi.org/10.1109/CMV.2003.1215005

[9] Kirley B.B. McCartt A.T. Braitman, K.A. and N.K. Chaudhary. 2008. Crashes of novice teenage drivers: Characteristics and contributing factors. *Journal of Safety Research* 39 (2008), 47–54. DOI: http://dx.doi.org/10.1016/j.jsr.2007.12.002

[10] Brake. 2018. UK road casualties: Brake the road safety charity. (dec 2018). https://www.brake.org.uk/facts-resources/1653-uk-road-casualties

[11] C. Brunsdon. 2001. The comap: exploring spatial pattern via conditional distributions. *Computers, Environment and Urban Systems* 25, 1 (2001), 53–68. DOI:http://dx.doi.org/10.1016/S0198-9715(00)00042-9

[12] Ward P. Bartle C. Clarke, D.D. and Truman. W. 2006. Young driver accidents in the UK: The influence of age, experience, and time of day. *Accident Analysis and Prevention* 38 (2006), 872–878. DOI: http://dx.doi.org/10.1016/j.aap.2006.02.013

[13] Higgs G. Brunsdon C. Corcoran, J. and A. Ware. 2007. The Use of Comaps to Explore the Spatial and Temporal Dynamics of Fire Incidents: A Case Study in South Wales, United Kingdom. *The Professional Geographer* 59, 4 (2007), 521–536. DOI: http://dx.doi.org/10.1111/j.1467-9272.2007.00639.x

[14] Guo F. Lee S. Antin J.F. Perez M. Buchanan-King M. Dingus, T.A. and J. Hankey. 2016. Driver crash risk factors and prevalence evaluation using naturalistic driving data. *Proceedings of the National Academy of Sciences of the USA.* 113, 10 (2016), 2636–2641. DOI: http://dx.doi.org/10.1073/pnas.1513271113

[15] R. Elvik. 2000. How much do road accidents cost the national economy? *Accident Analysis and Prevention* 32, 6 (2000), 849–851. DOI: http://dx.doi.org/10.1016/S0001-4575(00)00015-4

[16] S. Erdogan. 2009. Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research* 40, 5 (2009), 341–351. DOI: http://dx.doi.org/10.1016/j.jsr.2009.07.006

[17] Department for Transport. 2019a. Reported road casualties in Great Britain: 2018 annual report. *National Statistics* (2019), 47. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/834585/reported-road-casualties-annual-report-2018.pdf

[18] Department for Transport. 2019b. Road Safety Data. (sept 2019). https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

[19] D. Frigioni and L. Tarantino. 2003. Multiple zooming in geographic maps. *Data and Knowledge Engineering* (2003), 207–236. DOI: http://dx.doi.org/10.1016/S0169-023X(03)00060-0

[20] R. Goel. 2018. Modelling of road traffic fatalities in India. *Accident Analysis and Prevention* 112 (2018), 105–115. DOI: http://dx.doi.org/10.1016/j.aap.2017.12.019

[21] I.J. Good. 1983. The philosophy of exploratory data analysis. *Philosophy of Science* 50, 2 (1983), 283–295. https://www.jstor.org/stable/188015

[22] Welsh Government. 2019. Statistical First Release: Police recorded accidents, 2018. *Statistics for Wales* (2019), 23. https://gov.wales/sites/default/files/statistics-and-research/2019-06/police-recorded-road-accidents-2018.pdf

[23] E. Hauer and B.N. Persaud. 1987. How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effects of Warning Devices. *Transportation Research Record* 1114 (1987), 131–140. https://www.researchgate.net/publication/291841022_HOW_TO_ESTIMATE_THE_SAFETY_OF_RAIL-HIGHWAY_GRADE_CROSSINGS_AND_THE_SAFETY_EFFECTS_OF_WARNING_DEVICES

[24] Yang J.K. Huang, A.N. and F. Eklund. 2012. Analysis of car-pedestrian impact scenarios for the evaluation of a pedestrian sensor system based on the accident data from Sweden. *2nd International Conference of ESAR* (2012), 136–143. **https://bast.opus.hbz-nrw.de/opus45-bast/frontdoor/index/docld/396**

[25] N. Iliinsky and J. Steele. 2011. *Designing data visualizations: Representing informational Relationships.* O'Reilly Media.

[26] Children in Wales. 2019. Accident Prevention. (2019). **http://www.childreninwales.org.uk/our-work/accident-prevention/**

[27] Pasupathy R.K. Ivan, J.N. and P.J. Ossenbruggen. 1999. Differences in casuality factors for single and multi-vehicle crashes on two-lane roads. *Accident Analysis and Prevention* 31, 6 (1999), 695–704. DOI: **http://dx.doi.org/10.1016/S0001-4575(99)00030-5**

[28] Wang C. Ivan, J.N. and N.R. Bernardo. 2000. Explaining two-lane highway crash rates using land use and hourly exposure. *Accident Analysis and Prevention* 32, 6 (2000), 787–795. DOI: **http://dx.doi.org/10.1016/S0001-4575(99)00132-3**

[29] M.K. Janke. 1991. Accidents, mileage, and the exaggeration of risk. *Accident Analysis and Prevention* 23, 2-3 (1991), 183–188. DOI: **http://dx.doi.org/10.1016/0001-4575(91)90048-A**

[30] World Health Organization. 2018. *Global Status Report On Road Safety 2018: Summary.*

[31] G. Osborne and B. Turnbull. 2009. Enhancing Computer Forensics Investigation through Visualisation and Data Exploitation. In *2009 International Conference on Availability, Reliability and Security*. 1012–1017. DOI: **http://dx.doi.org/10.1109/ARES.2009.120**

[32] Xia J.C. Plug, C. and C. Caulfield. 2011. How to Estimate the Safety of Rail-Highway Grade Crossings and the Safety Effects of Warning Devices. *Accident Analysis and Prevention* 43 (2011), 1937–1946. DOI: **http://dx.doi.org/10.1016/j.aap.2011.05.007**

[33] L.P. Rieber. 1995. A historical review of visualization in human cognition. *Educational Technology Research and Development* 43, 1 (1995), 45–56. DOI: **http://dx.doi.org/10.1007/BF02300481**

[34] Regev S. Moutari S. Rolison, J.J. and A. Feeney. 2018. What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records. *Accident Analysis and Prevention* 115 (2018), 11–24. DOI: **http://dx.doi.org/10.1016/j.aap.2018.02.025**

[35] Athanasios Salamanis, Ilias Kalamaras, Alexandros Zamichos, Anastasios Drosou, Dionysios Kehagias, Stavros Papadopoulos, Dimitrios Tzovaras, and Georgios Margaritis. 2017. An Interactive Visual Analytics Platform for Smart Intelligent Transportation Systems Management. *IEEE Transactions on Intelligent Transportation Systems* PP (07 2017). DOI: **http://dx.doi.org/10.1109/TITS.2017.2727143**

[36] N. Saunier and T. Sayed. 2007. Automated Analysis of Road Safety with Video Data. *Journal of the Transportation Research Board.* (2007), 57–64. DOI: **http://dx.doi.org/10.3141/2019-08**

[37] Hermans E. Brijs T. Wets G. Shen, Y. and K. Vanhoof. 2012. Road safety risk evaluation and target setting using data envelopment analysis and its extensions. *Accident Analysis and Prevention.* (2012), 430–441. DOI:**http://dx.doi.org/10.1016/j.aap.2012.02.020**

[38] B. Shneiderman. 1996. The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*. 336–343. DOI: **http://dx.doi.org/10.1109/VL.1996.545307**

[39] Mackenzie J.R.R. Dutschke J.K. Baldock M.R.J. Raftery S.J. Thompson, J.P. and J. Wall. 2018. A trial of retrofitted advisory collision avoidance technology in government fleet vehicles. *Accident Analysis and Prevention* 115 (2018), 34–40. DOI: **http://dx.doi.org/10.1016/j.aap.2018.02.026**

[40] J.W. Tukey and P.A. Tukey. 1988. Computer Graphics and exploratory data analysis: An introduction. *The Collected Works of John W. Tukey: Graphics.* 5 (1988), 419–431.

[41] P. Pramada VALLI. 2005. ROAD ACCIDENT MODELS FOR LARGE METROPOLITAN CITIES OF INDIA. *IATSS Research* 29, 1 (2005), 57 – 65. DOI: **http://dx.doi.org/https://doi.org/10.1016/S0386-1112(14)60119-9**

[42] Z. Wang, M. Lu, X. Yuan, J. Zhang, and H. v. d. Wetering. 2013. Visual Traffic Jam Analysis Based on Trajectory Data. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (Dec 2013), 2159–2168. DOI:**http://dx.doi.org/10.1109/TVCG.2013.228**

[43] A.M. Williamson and A.M. Feyer. 1995. Causes of accidents and the time of day. *Journal of Work and Stress.* 9, 2 (1995), 158–164. DOI: **http://dx.doi.org/10.1080/02678379508256550**

[44] K. Yamaba and Y. Miyake. 1993. Color character recognition method based on human perception. *Optical Engineering.* 32, 1 (1993). DOI: **http://dx.doi.org/10.1117/12.60072**