CAR CRASH ANALYTICS

Aquilina Mattia
Sapienza University of Rome
aquilina.1921153@studenti.uniroma1.it
https://github.com/Mattia-Aquilina

Zanchetta Luca
Sapienza University of Rome
zanchetta.1848878@studenti.uniroma1.it
https://github.com/luca-zanchetta

ABSTRACT

Road accidents pose significant risks to public safety and incur substantial economic costs. Analyzing accident data is crucial for understanding the underlying causes and patterns to devise effective interventions. This paper presents **Car Crash Analytics**, a visualization tool designed to analyze car accident data for promoting driving safety, particularly aimed at driving school instructors. The tool offers an interactive interface featuring geographical map, scatterplot, heatmap, and parallel coordinates charts to identify patterns, trends, and potential risk factors associated with road accidents. Through data pre-processing and visualization techniques, users can explore insights such as safer areas and conditions for driving lessons. The tool's capabilities empower instructors to make informed decisions and contribute to reducing road accidents through evidence-based strategies. While offering valuable insights, the tool also presents opportunities for further enhancements and collaboration with stakeholders to advance road safety initiatives.

***K*eywords** Visualization Tool · Visual Analytics · Driving Safety · Road Safety · Driving School

1 Introduction

Analyzing road accidents data is crucial for promoting driving safety, as it offers invaluable insights into the causes, trends, and patterns underlying these incidents. Beyond injuries and fatalities, road accidents impose substantial economic costs on society and disrupt communities. By delving into accident data, we can identify recurring factors such as speeding, distracted driving, and road conditions that contribute to accidents, allowing for targeted interventions and policy measures. Moreover, understanding accident data enables us to design effective public awareness campaigns and educational initiatives aimed at fostering safer driving behaviors. Ultimately, continuous analysis of road accidents data facilitates the development of evidence-based strategies to prevent accidents and mitigate their impact, underscoring its significance in safeguarding lives and enhancing road safety for all the people.

The **Car Accident Dataset** provided by Kaggle [1], for instance, provides detailed records of road accidents that occurred during January 2021 in Kensington and Chelsea. It includes information such as the accident date, day of the week, junction control, accident severity, geographical coordinates, lighting and weather conditions, vehicle details, and more. The data is valuable for analyzing and understanding the factors contributing to road accidents in this urban area, aiding in the development of strategies for improved road safety. For the purposes of this paper, we have considered a subset of this data consisting of around 3K tuples, polished in a proper way according to our analytics objectives. See 3 for further details.

For the purposes we have discussed, we introduce **Car Crash Analytics**, a visualization tool whose aim is to provide an interactive and user-friendly interface through which users are able to discover causes, trends and patterns of car accidents, with the final purpose of improving driving safety. In particular, this tool has been designed for *driving school instructors*, having the goal of finding the best places and conditions under which to perform driving lessons.

2 Related Works

Research into car accidents encompasses various aspects including causes, consequences, prevention measures, and mitigation strategies. In this section, we review relevant tools in the literature focusing on key themes within car accident studies.

ArcGIS Online [2] is a visualization tool providing a comparison between crash severity, crash manners and non-motorist agents involved in the crash by the means of a map view and several bar charts. The user is able to filter the data in several ways, and the interaction is limited to the data displayed in the current position of the user on the map view.

The **Iowa Department of Transportation** [3] provides a visualization tool that consists of a map view containing all the data about car accidents, without any severity discrimination. The tool provides the possibility to change the map view, and allows users to create custom charts for visualizing the data. It also allows users to generate reports about the filtered data.

Power BI Report [4] is a tool provided by Microsoft that allows to visualize data about car accident by the means of several charts, showing the data filtered by year, by month, by district, by type and by route. It also offers a map view provided by Microsoft Bing, that shows only representative samples, and not all the available data. Users can navigate throughout the tool for gaining insights; this is possible thanks to the multiple screens the tool is offering.

All the above-mentioned visualization tools were using different data with respect to the tool proposed by this paper. Moreover, as we will further explain in section 3, none of the above-mentioned visualization tools provide the possibility to brush the data to analyze clusters of car accidents, in order to identify their causes. We think that this possibility is crucial for identifying car accident causes and therefore insights that help improving driving and road safety.

The only relevant available study on the same dataset we have used, the *Car Accident Dataset* [1], is the one conducted by Muhamed S. [5]. It is not a visualization tool; therefore, there is no user interaction. However, it provides interesting visualizations like for instance a bar chart showing the relation between the number of crashes and the road surface conditions; another visualization showing the percentage of accidents by the day of the week, and another visualization showing the accident count by speed limit.

3 Methodology

The **Car Crash Analytics** visualization tool is a React-native application providing a dashboard that allows users to gain insights about the data by interacting with the system itself. The whole application has been coded by using several programming languages, such as JavaScript (and, in particular, the React.JS framework), TypeScript and Python. In what follows, we will explain all the steps involved in data preprocessing, including used dimensionality reduction techniques. Moreover, we will also detail the visualization and interaction techniques we have employed in the application, in order to improve as much as possible the user experience.

3.1 Data Preprocessing

In this section, we explain the steps involved in data preprocessing, including the used dimensionality reduction techniques. We remind the reader that we have considered the *Car Accident Dataset* [1] from Kaggle; this dataset was providing over 300.000 tuples of 21 columns. However, since the whole dataset was too big for our application, we have decided to consider a smaller subset of only 2900 rows. This smaller subset perfectly fits the needs of the application, avoiding crashes and the unresponsiveness of the UI.

We have properly polished this dataset by removing the columns *Local Authority (District)*, *Carriageway Hazards* and *Police Force*, since we consider them unnecessary from the point of view of a driving school instructor. Additionally, we have reinterpreted the accident indexes, since those proposed in the dataset were big unreadable integers; in particular, we have simply enumerated all the rows starting from 0. Then, we have grouped all the time occurrences of the car accidents into hourly intervals: 00:00 - 02:59, 03:00 - 05:59, 06:00 - 08:59, and so on, until the last interval 21:00 - 23:59. After that, we have grouped in sets all the values of all the categorical attributes, with the aim of visualizing them in a parallel coordinates chart (see section 3.2.4), in order to let the user have a better understanding of the causes of the car accidents under analysis.

Finally, we have applied a dimensionality reduction technique in order to show six attributes in a scatterplot chart (see section 3.2.2). In particular, we have used the **t-SNE** dimensionality reduction technique applied to the following attributes of the data: *Number of Causalties*, *Number of involved Vehicles*, *Speed limit*, *Latitude*, *Longitude* and *Accident Severity*. We have applied t-SNE to these attributes because they were the only numerical attributes that were present in

the data, with the aim of finding some clusters that could help the intended user to find some insights about the causes of potentially fatal accidents. Additionally, we have chosen to apply t-SNE instead of PCA or MDS because, although it introduces the possibility to have false positives and false negatives, it amplifies the separation between clusters of the data, that are arranged on the 2D space of the scatterplot chart by iterating and minimizing a stress function. Since our aim was to find clusters of the data in order to have a better understanding of the causes of car accidents, this solution was the one that worked the best for our purposes.

3.2 Employed Visualizations

In this section, we will detail the visualization and interaction techniques we have employed in the system. All the charts we are going to discuss are interactive: that is, each chart offers to the user the possibility to interact with the displayed data. The result of each interaction is reflected in all the other charts. For instance, if the user selects a cluster displayed on the scatterplot, the selected data will be shown in all the other charts, along with the applied filters.

Our application consists of a dashboard showing the filters section and four charts: a *geographical map*, a *scatterplot chart*, a *heatmap chart* and a *parallel coordinates chart*. A screen of our dashboard is reported in figure 1.

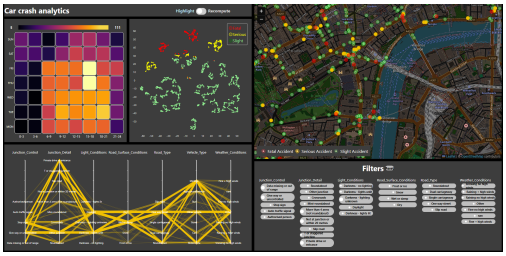

Figure 1: The dashboard of Car Crash Analytics.

3.2.1 Geographical Map

The geographical map contains all the car accidents of the considered subset of the original data. We have placed the points onto the map by considering the latitude and the longitude of each car accident; this information was available in the dataset. By having a look at figure 1, we can see that points are colored according to the severity of the single accidents: fatal accidents are red, serious accidents are yellow and slight accidents are green. The user can select a specific area by brushing onto the screen, and the corresponding selected points will be visualized in all the other charts, along with the activated filters.

3.2.2 Scatterplot Chart

The scatterplot chart contains data points resulting from the application of the t-SNE dimensionality reduction technique. As in the case of the geographical map, also the scatterplot points are colored according to the corresponding accident severity. The user can select data points by brushing, zooming and panning onto the chart; obviously, the corresponding selected points will be visualized in all the other charts, along with the activated filters.

A peculiarity of our scatterplot chart is that it has two visualization modes: *highlight* and *recompute*. Selection of multiple modes simultaneously is not permitted; users may choose only one at a time. Within the *highlight* mode, the t-SNE algorithm is executed only at the beginning of the interaction, and the filtered data are highlighted onto the chart. In simple words, there is no scatterplot data loss upon filtering the data. Conversely, within the *recompute* mode, the t-SNE algorithm is executed each time the user introduces new filters. This functionality is designed to assist users in uncovering potentially obscured clusters within the dataset.

3.2.3 Heatmap Chart

The heatmap chart contains data about accidents in specific days of the week and specific time intervals. In particular, each square of the heatmap contains the total number of accidents of the considered portion of the data in a particular time interval of a particular day of the week. For instance, let us consider the top leftmost square of the heatmap shown in figure 1: if the user goes onto this square with its mouse, (s)he will get the total number of car accidents that happened on Sundays, between 12:00 AM and 03:00 AM. All the squares of the heatmap are colored according to the Inferno palette, which provides the user with a sense of increasing brightness, that is obviously related to the increasing

number of car accidents we are considering. The more bright is a specific square, the more accidents have occurred during that particular day of the week in that particular time interval.

3.2.4 Parallel Coordinates Chart

The parallel coordinates chart allows to visualize some of the categorical attributes of the data, in order to find insights about the causes of the accidents that we are taking into account. Indeed, when the user has applied several filters on the data, (s)he may want to discover the causes of the selected accidents, in order to find insights about the safest places or conditions under which to perform driving lessons. The user is able to interact with the parallel coordinates chart by brushing onto the axes, each representing one categorical attribute. The categorical attributes of the data that are shown by the parallel coordinates chart are the following: *Junction Control*, *Junction Detail*, *Road Surface Conditions*, *Light Conditions*, *Road Type*, *Type of Vehicle involved* and *Weather Conditions*.

4 Discovered Insights

Let us follow an execution workflow of the visualization tool, in order to find some insights about the data.

4.1 Most safe area

First of all, we can have a look at the geographical map chart depicted in figure 2.

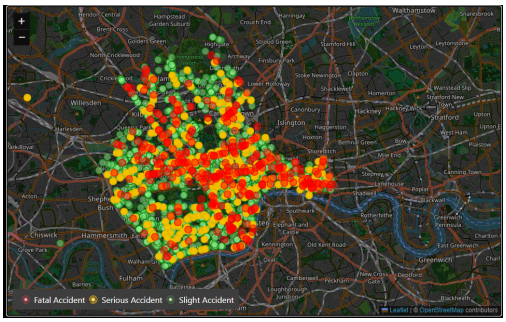

Figure 2: Geographical map insights.

As we can see from the above picture, the majority of the fatal accidents are located in the center of the city; therefore, the intended user should avoid to perform driving lessons there. Instead, the user should prefer safer areas such as the north one or the south-west one.

4.2 Daylight driving lessons

Now, we would like to know the best area and the best conditions under which to perform driving lessons in the morning or in the afternoon. For this purpose, we filter the data on the heatmap by considering as working days Monday, Tuesday, Wednesday, Thursday and Friday, and as working time intervals 09:00 - 12:00, 12:00 - 15:00 and 15:00 - 18:00. Since we are excluding the evening from our analysis, we also set the Daylight filter, thus obtaining a situation like the one depicted in figure 3.

Figure 3: Daylight insights.

By having a look at the geographical map chart of figure 3, we can see that the most safe area is the south-west one, since the majority of the accidents are those of slight severity. Therefore, by selecting this area, we obtain a situation like the one depicted in figure 4.

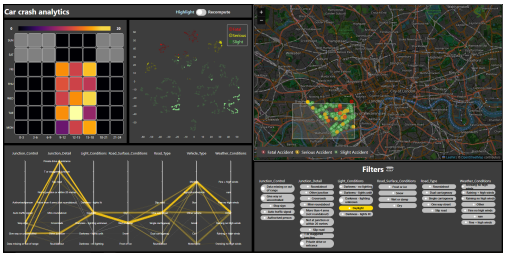

Figure 4: South-west area insights.

As we can see from the shown parallel coordinates chart, we have to be careful in the following situations:

- When the road surface is wet;
- When we encounter a minibus or a motorcycle;
- When the road has a T junction or is intersected with a private drive or entrance;
- When there are stop signs or give way signs.

Moreover, if we have a look at the heatmap chart, it seems that the most safe time interval in which to perform driving lessons is the 9:00 - 12:00 interval, while the most safe day of the week is Monday.

4.3 Evening driving lessons

Now, we would like to know the best area and the best conditions under which to perform driving lessons in the evening. For this purpose, we filter the data on the heatmap by considering as working days Monday, Tuesday, Wednesday, Thursday and Friday, and as working time intervals 15:00 - 18:00 and 18:00 - 21:00. Since we are excluding the daylight from our analysis, we also set all the available Darkness filters, thus obtaining a situation like the one depicted in figure 5.

Figure 5: Evening insights.

By having a look at the geographical map chart of figure 5, we can see that the most safe area is the north one, since the majority of the accidents are those of slight severity. Therefore, by selecting this area, we obtain a situation like the one depicted in figure 6.

Figure 6: North area insights.

As we can see from the shown parallel coordinates chart, we have to be careful in more or less the same situations we discovered for the daylight case, except for the fact that we must be careful when the road lights appears to be off.

Moreover, if we have a look at the heatmap chart, it seems that the most safe time interval in which to perform driving lessons is the 15:00 - 18:00 interval, while the most safe days of the week are Monday and Thursday.

5 Conclusions

The Car Crash Analytics visualization tool has provided valuable insights into road accident data, facilitating informed decision-making for driving instructors and promoting driving safety. Through interactive visualization techniques such as geographical map, scatterplot, heatmap and parallel coordinates charts, users can explore patterns, trends, and potential risk factors associated with road accidents.

Key findings from the analysis include identifying *safer areas* and *conditions* for conducting driving lessons, such as avoiding central city areas during peak traffic times and prioritizing daylight hours for instruction. By leveraging the tool's capabilities, driving instructors can make data-driven decisions to enhance the safety of their lessons and ultimately contribute to reducing road accidents.

The project underlines the importance of utilizing visualization tools for analyzing road accident data, as they provide a comprehensive understanding of underlying factors and enable stakeholders to develop targeted interventions for improving driving safety.

While the Car Crash Analytics tool offers valuable insights, it is not without limitations. Future research should focus on addressing these limitations, for instance by improving the limited brushing on the geographical map view, or considering the whole *Car Accident Dataset* for gaining newly and potentially more useful insights. Moving forward, it is recommended to expand the tool's functionalities, incorporate real-time data updates, and collaborate with relevant stakeholders to implement evidence-based strategies for road safety.

References

[1] Kaggle. Car Accident Dataset. `https://www.kaggle.com/datasets/nextmillionaire/car-accident-dataset/data`. Accessed: February 21, 2024.

[2] Arcgis online. `https://experience.arcgis.com/experience/1911f992cabc484a98f64e7c36c2b262/`. Accessed: February 23, 2024.

[3] Iowa Department of Transportation. `https://icat.iowadot.gov/`. Accessed: February 23, 2024.

[4] Power BI Report. `https://app.powerbigov.us/view?r=eyJrIjoiMjhlZjFhZDAtNTljMC00MDA1LWEyOTMtYWYwM2NiMmRiMm` Accessed: February 23, 2024.

[5] NextMillionaire. Car accident full data analysis. `https://www.kaggle.com/code/nextmillionaire/car-accident-full-data-analysis`. Accessed: February 23, 2024.