Countrywide traffic accident dataset visualization

Remondini Leonardo, Molon Mattia

1 Introduction

As of today, cars are the most used vehicle, and car accidents are one of the most common death cause worldwide. If we examine the USA trend only, we would find that more than 6 million car accidents take place every year, killing more than 90 people per day. The improvement of traffic restoration and the prevention of those accidents has to be one of the main activities of local authorities and police. However, sometimes it can be hard to know exactly where and when accidents are likely to occur, and not always local authorities efficiently succeed in traffic restoration. Is there something we can do to improve human intervention, prevent car crashes, and consequently decrease the occurrence of those?

During the last three years, the USA government collected an enormous amount of data about these misfortunes and indeed it is trying to gather information and take action. Here, is where our project would be helpful. Within this paper, we will introduce a tool thought for the US local police or public administration that could help them to analyze the amount of car accident data available, and retrieve unexpected outcomes. We want to facilitate the way the user gets access to real information that comes from a giant amount of raw data. Our framework aims to display a simple-to-read overview of car accidents trend and also to find, if there, a correlation among weather conditions, road structure, place, and time of all car crashes occurred. In this way, we will be able to detect "dangerous zones" that could be safely restructured, or even cities that struggle the most with traffic restoration. Therefore, we aim to raise awareness of what the data seems to tell us: where and when to act.

2 Problem Description and Task Analysis

The US government provides every year a bunch of different statistics related to car accidents. As shown in the ones of 2013 [3], car crashes are one of the main causes of injuries in the USA. 6 million misfortunes occur every year, 50% of which end up in injuries with more than 2 million of permanent ones. There is clearly a need to somehow predict where and when the most dangerous accidents take place and how the road setting could influence the accident characteristic to develop safer roads. Moreover, it is important to understand whether the weather condition influences the total number of accidents and how to react. Furthermore, those accidents have consequences on the traffic flow. Indeed it can cause lots of discomfort to citizens and police forces. As we explained in the previous paragraph, the consequences of car crashes can be devastating if not handled in the right way. The restoration of the traffic normal traffic flow has to be efficient and fast. But not all the cities can claim to have efficient teams. How can we do to detect those cities that most need help in managing car accident consequences? The framework we will introduce aims to provide a public tool that can visualize and summarize car accident data, in order to facilitate public authorities to deal with such problems.

• *Remondini Leonardo and Molon Mattia are with Eindhoven University of Technology. E-mail: {l.remondini, m.molon}@student.tue.nl.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

2.1 Dataset

The data [1] that we want to analyze and visualize consists of a countrywide traffic accident dataset. It covers 49 states of the United States and it contains data detected using several data providers such as Bing and MapQuest, including different APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the IJS and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. The data has been collected from February 2016 to December 2019 and currently, it's composed of around 3 million accident records. The dataset is structured as a CSV table. Every row represents a different car accident, and the columns its characteristics. The attributes recorded for every tuple consist in: the weather and environment condition at the time of the accident, its exact geographic coordinates (latitude and longitude), the type of street, the severity of the accident, its description, how many miles of queue it provoked and for how long the condition of the traffic remained altered. All the present numeric data uses American scales of measurements such as Farheneit, Miles and inches.

2.2 Tasks analysis

The analysis of this data can bring interesting insights about the accidents in the US, which can be used by American police forces and municipalities to understand better which are the general trends of the car crashes in the different countries, and how to handle better these misfortunes. More precisely, we believe that throughout our visualization tool we will be able to:

- Understand the general trends of car accidents in America. How many misfortunes are happening every year in every geographical area of the country (cities, counties or states) and how severe are them.
- Understand which is the topology of the most common roads related to car crashes and how the topology are related one another.
- Understand if there is a correlation between the weather environment conditions (presence of clouds, rain, speed of the wind, humidity, etc.) and the car crashes characteristics
- Understand how these accidents are handled by the authorities, how much it usually takes to restore the normal traffic flow and which are the geographical areas that struggle the most.
- Make a forecast of the future expected trends of accidents using statistical techniques.

2.3 Preproccessing

3 million tuples reported in 4 years are a gigantic mole of records, they consist of an average of $\approx$ 62,500 per month and $\approx$ 2,000 per day. Therefore, the dataset is extremely likely to contain multiple outliers, null values, and redundant information. Indeed, it necessitates a pre-processing process in order to be correctly analyzed.

The first point to notice is that the data has been collected by two main sources, MapQuest and Bing. Both of the actors started collecting data from the same period in time (January 2016), but from different locations. Bing has always gathered information from the whole states, on the other hand, MapQuest started doing so only after August 2017. Before that date, the only locations covered by MapQuest were 26 out of

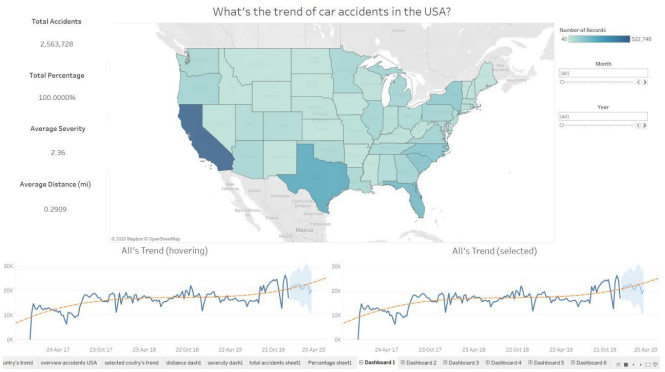

Fig. 1. Dashboard 1: What's the trend of car accidents in the USA?

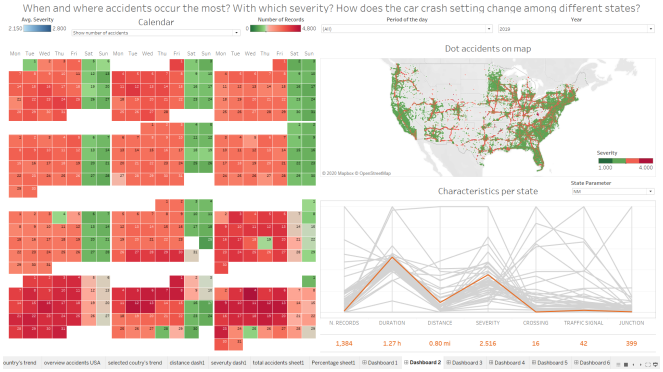

Fig. 2. Dashboard 2: When and where accidents occur the most? With which severity? How does the car crash setting change among different states?

the 49 states present in the dataset. This scenario precludes us from the possibility to use data before august 2017 whilst making investigations and comparisons between states. Indeed, this misbalanced scenario is likely to wrongly influence the computed statistics and forecasts. For this reason, we decided to exclude from our visualizations all the data that concern these periods in time, taking into consideration only the years 2017-18-19.

Another important aspect is the difference in the severity of the car crashes between the sources. In the description of the dataset, this attribute is described as a numeric value ranged between 1 and 4 where the value 1 stays for "the accident caused a short delay on traffic" and 4 stays for "the accident caused a long delay and discomfort". The two sources are supposed to use the same scale but data mine the information reveals how this is not the case. Indeed, Bing's severity indicator ranges between 2 and 4 and has the majority ($\approx 40\%$) of its records classified as level 4 severities. Considering that this value should be associated only with extremely unlikely events, it seems reasonable to us to doubt their reliability. Moreover, conduct topic research does not give satisfiable results on the reasons behind these differences between the sources. Hence, we decided to exclude Bing's data for severity analyses.

Moving on, for every attribute in the dataset, clear outliers are present. For instance, accident durations which last 256 hours or queues 333 miles long. Hence, for each attribute field, we decided to take into account only the attribute values which lay inside of the range $[0, p_{99}]$, where p_{99} is the 99th percentile of the attribute values taken into consideration. This decision gives reasonable results such as the maximum lasting time of 18 hours and the maximum traffic queue of 11 miles.

3 Visualization Design

To proceed with the visualization of the data we chose to exploit Tableau [2] as a Visualization framework. Our project consists of six different dashboards that interact with more than ten visualization techniques. Each of these dashboards investigates and answers a different question. By doing so, the user won't have to change multiple dashboards to answer a question, but he will have all the instruments to do so in a single dashboard.

3.1 Visualization Choices

To give the user the best possible user experience, we decided to visualize the data parameters as follows:

- To visualize the number of accidents among states, we decided to exploit a map and color every state based on how many accidents have occurred there. The color palette used ranges from light blue to dark blue.

- To visualize trends over time we exploit a blue simple line plot fitted by a scattered line that captures the global trend. On the other hand, to visualize trend comparisons, we exploited PCPs (parallel coordinate plots) and area charts. Every attribute of the charts is colored differently.

- To examine when accidents occur the most, we developed a calendar heat map and colored every day based on the average severity or the average number of accidents (customizable by the user). On the other hand, to visualize where accidents occur with the highest severity rate, we displayed every accident on a map as a point, as big as the severity rate they describe. Moreover, every point is colored from green to red, with the red capturing high rates of severity and green capturing the low rates of severity. The severity palette has been so chosen for psychological reasons. The green always describes a good situation, whilst the red always describes the bad one. This reasoning has been applied also to visualize distance and duration on the map.

- To visualize correlations, we chose to exploit a butterfly bar chart. In this way, the user will be able to instantly recognize negative correlation (as they are displayed on the left side of the cart) from the positive ones (as they are displayed on the right side of the cart). Correlations are colored with a green-to-red palette for psychological reasons.

- Road settings affect the number of accidents in every city differently. To capture this difference we decided to use a simple bar chart, where each bar represents a city and is colored based on its specific road settings subdivision. This feature will give the user an overview of the cities most affected by a certain road setting.

3.2 Dashboards design development

The project opens with a dashboard (Figure 1) that presents an overview of the car accidents trend in America. More precisely our data covers 49 out of the 51 States that form the United States of America, excluding Hawaii and Alaska. The aim is to quantify and compare the total amount of accidents that occurred between January 2017 and December 2019. There are four main visualizations on the first dashboard. The first and biggest plot is an interactive map that shows, by the usage of a color scale, the total amount of accidents that occurred within each state during a certain period of time. By default, the time-slot is set from 2017 to 2019, but the user can choose himself the month and the year to investigate by using the two filters located on the right side of the map. These filters are applied to the whole dashboard. On the left side of the map, there are instead four fields that present some trend details about the whole country or a specific state if required. In fact, the user can select or hover a specific state on the map to highlight it and visualize the trend details about that specific state. There are displayed (in numbers) the total amount of accidents, the percentage that it describes on the whole dataset, the

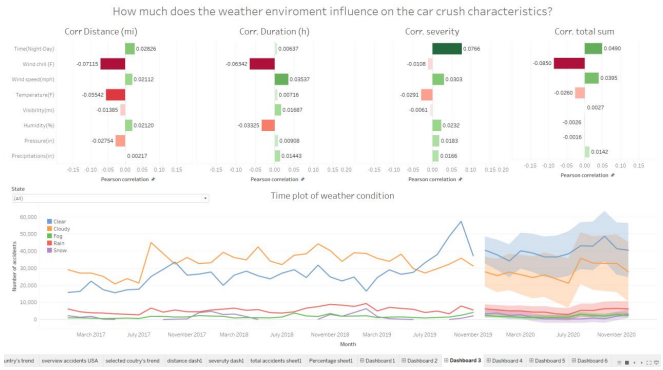

Fig. 3. Dashboard 3: How much does the weather environment influence on the car crush characteristics?

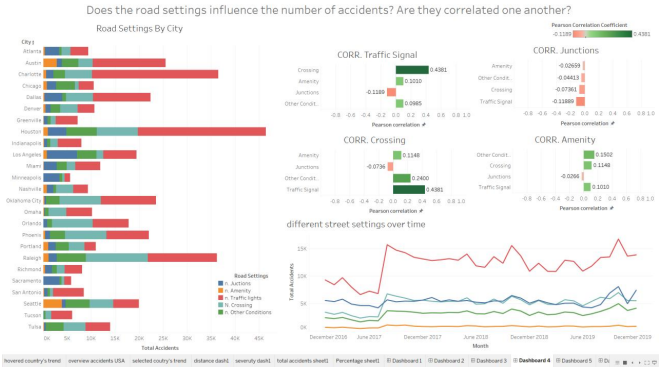

Fig. 4. Dashboard 4: Does the road settings influence the number of accidents? Are they correlated one another?

average severity of those accidents, and the average distance (in miles) that those accidents have caused. On the bottom of the dashboard, two more identical plots describe the time trend of the number of accidents, within the time-slot chosen, per week number. The plots differ on what trend they are showing. The first line chart shows the time trend of accidents of the hovered state and the second one shows the time trend of accidents of the selected one. Both of them describe as default the time trend of the whole country. As a result, the user will be able to compare the time trend of a state with one of another state or of the whole country itself. Moreover, if the user selects a specific week number on the parallel plot, he will focus the whole dashboard on the analysis of that specific week. The plots also present the amount of accidents expected in the future, based on the data mining techniques used by Tableau.

In the second dashboard (Figure 2), we wanted to investigate in-depth the time and spatial location of the accidents, whilst showing an overview of the most frequent characteristics of the accidents per state. To do so, we developed three different plots: one that describes when these injuries occur the most, one that describes where they occur the most, and one that shows how the most important and frequent car crash characteristics differ among all states . The time location is displayed by a calendar heat-map located on the left side of the dashboard and it shows the total amount of accidents performed on each calendar day during the year selected. By default, the year is 2018, but the user can choose whatever year he's interested to examine (from 2017 to 2019). The user can also choose, using a filter, to visualize the average severity of the accidents per day, instead of the trend of the total amount of accidents. Moreover, by using another filter, the user can choose whether to display the accidents that occurred during the day, the night, or both. All these filters can be found at the top of the dashboard and affect the whole dashboard. Last but not least, if the user selects a calendar day, all the plots of the dashboard will focus on the analysis of that specific day. By exploring this calendar, the user will be able to recognize which are the days with more car crashes and if they are correlated with specific periods of the year. To answer the question "Where?", we chose to visualize each car crash location on a map as a point colored by its severity value. From this plot, the user will be able to locate potential "dangerous spots" where many car crashes occurred with a high severity rate. The map interacts with all the filters we described before. Finally, to overview how the most frequent characteristic of the accidents ranges among the states, we chose to visualize a PCP (parallel coordinate plot), where each column represents a different crash characteristic. The user will be able to visualize which states deviate from the normal trend and could be the subject of further analyses. Moreover, the user can select a specific state to highlight his path on the plot, and filter all the data visualized on the dashboard to be focused on that state.

In the third dashboard (Figure 3), we explored how much the weather could influence three of the main characteristics of a car crash. We focused on how some parameters like light, wind speed, visibility (in miles), precipitations, and more could be correlated with the type of accident that occurs. That means if the environment could raise parameters as severity, duration, distance, and a combination of those (the sum of the parameters scaled by their maximum). To do so, we created, at the top of the dashboard, four different plots that show the Pearson correlations coefficient between weather parameters and car crash characteristics. We decided these plots to be two-sides horizontal bars as the user will know immediately which car crash parameters are more likely to increase or decrease when occurred in a certain environment. The bars' color ranges from green to red, with green capturing a positive correlation between the two parameters, and with red capturing the negative ones. Moreover, at the bottom of the dashboard, we added the time trend (in weeks) of the number of accidents occurred with more general weather conditions as cloudy, snowing, clear, foggy and rainy. By selecting a particular week within the plot, the user will filter all the Pearson correlation coefficient to be focused on the week selected, related to the weather condition chose. The user also can find a city filter that focuses the whole dashboard on the analysis of a particular city.

In the fourth dashboard (Figure 4), we explored another type of correlation. We studied how much the road settings and the proximity to certain residential areas could weight on the total amount of accidents occurred, and if a combination of those could increase the probability of an accident to take place. In particular, we examined the three most frequent road settings (crossing, traffic light, junction), the most frequent type of residential area affected (amenity zones) and the sum of all the minor road settings (bump, turning loops, roundabout, railway ext.). On the left, we can find a colored stacked bar chart, and it describes the number of accidents occurred within a certain road environment among the top 15 cities with the highest amount of accidents. The bars are divided by colors that capture the range of the total accidents affected by a particular road setting, or amenity zone proximity. In this way, the user will be able to identify where some road settings affect the most and how they are distributed among all city car crashes. If the user selects a particular city the dashboard will focus on the analysis of the city selected. On the right side of the dashboard, there are two main sections. On the top right side, we found 4 different plots that quantify the correlation that each road setting has on the others. For each main setting (amenity zone proximity, traffic light, crossing, and junction) there are visualized four Pearson correlation coefficient that captures all the combination of road settings, plus the relation with other road settings. The time and zone referred to these correlation coefficients can be filtered by the other sections. The user will be able to understand which are the combination of road settings that most cause accidents. Finally, on the bottom right side of the dashboard, there is displayed a plot that describes the time trend of accidents occurred within a certain road setting or near amenities. If

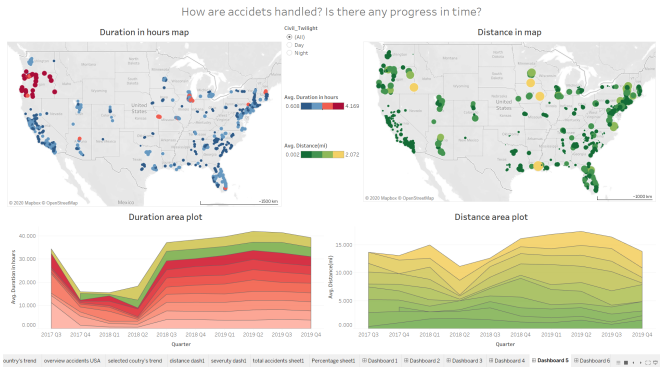

Fig. 5. Dashboard 5: How are accidents handled? Is there any progress in time?

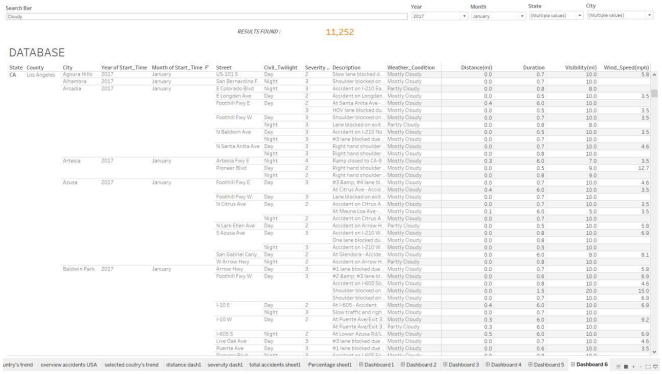

Fig. 6. Dashboard 6: Database research

the user selects a particular month, the dashboard will focus on the analysis of the selected month.

In the fifth dashboard (Figure 5), we studied how the accidents are handled by the authority. The user will be able to understand which parts of the USA have difficulties in handling the injuries and need help to improve the service. To help the user to perform this task, we placed at the top of the dashboard two map plots, which show how the average duration of an accident and the average length of the road extent, affected by the accident, respectively change across the country, per city. Each car crash is displayed as a point, as big as the duration or the distance that has caused. Under the maps, there are two other similar plots. Each of those refers to the map that they have at the top and wants to highlight the time trend of the variables "distance" and "duration" of the 10 cities with the highest values. The user will be able to recognize the cities that most need external help and observe if the trend has changed during the years. The user will be able also to decide whether to visualize the trend of accidents occurred at night, at daylight, or both (for all plots). If the user hovers or selects a particular city, its trend on the other plots will be highlighted.

Finally, in the sixth and last dashboard (Figure 6), we wanted to let the user be free to dig into the dataset (showed as a table) and search for further information or particular fields. The dashboard is a simple search bar that shows all the results retrieved form the research in the dataset. We added the possibility to filter the search by year, month, state and city.

4 Use Cases

As broadly outlined in Section 3, our visualization tool consists of 6 dashboards and more than 10 different visualization techniques. Each dashboard aims to investigate an aspect of the dataset whilst exploring the data. We now go through all the dashboards one by one and explain which information they show and which insights can be extrapolated from each one of them.

"Understand the general trends of car accidents in America. How many misfortunes are happening every year in the different geographical areas of the country (cities, counties or states)? How much severe are them?"

These questions are relatively easy to answer while looking at the first dashboard (Figure 1). The information that should immediately pop up to the eyes is the strong difference in the color of three states compared to all the others. These states are California, Texas, and Florida and indeed represent the geographical areas in which the car accidents happen the most. Respectively, California detains the record of 522,748 tuples in 3 years, nearly 20% of the entire dataset and more than twice the national average, Texas holds second place with 248,297, and Florida follows with 186,713. Overall, it is clear how the number of accidents is distributed mainly along the two coasts (which are colored by a darker blue) and how in the center are less on average (lighter shades of blue). This distribution should raise some questions such as: Why accidents in California are so frequent with respect to other places in the USA? Why is it more common to register accidents on coasts rather than in the central states? Is it due to the street/weather conditions? Is it due to how the data was collected?

Moving to the bottom charts, we can observe how the accident trends change over time among different parts of America. At first, looking at the overall plot, it can be noticed how the trend seems to be increasing. The regression line remains approximately stable during the years 2017 and 2018 but starts increasing in value during 2019. Furthermore, the forecast computed on the data reveals how the predictions for future months have a positive trend. Hovering over the different countries and using the comparison feature, we can also notice how the number of accidents remains stable along the east coast during the three years whilst it increases on the west coast. The most significant increment takes place in the state of California. It goes from a mean average of 3000 misfortunes per week to a mean average of 7000. This data is suspicious and suggests further analysis.

"Understand in a more precise way where and when accidents occur the most, with which severity and how does the car crash setting changes among different states"

The second dashboard (Figure 2) comes in our help to shed light on these topics. The dotted map in the top-right corner shows a really interesting insight into the exact positions in which the accidents occur and with which severity. The user can notice how the dots reconstruct almost perfectly the street network of the states. The highways are highlighted in red color, and all the cities and countrysides are characterized by a green one. The color of the dots represents the severity of the accidents, hence, it suggests how the ones occurring in the highways are usually far more severe than the one happening in other types of streets. This is probably due to the velocity with which cars are used to travel in these types of streets, and consequently in the more dangerous scenarios that car crashes produce. Once again, the map suggests that the accidents are mainly concentrated on the two coasts and less likely to happen in the middle states.

On the left side of the dashboard, the heat map calendar gives a clear view of how the accidents are distributed during the different days of the years. Immediately, it is noticeable how during weekends a smaller number of accidents are happening compared to the working days but, on the other hand, they are usually more severe. The first reasoning behind this feedback can bring to the following conclusion: during the working days, the majority of people use the car to reach workplaces and offices. This flow of people causes a lot of traffic which proceeds at a slow pace. Therefore, it is likely to record low severity accidents like little collisions. On the other hand, during weekends, people drive faster and cause severe accidents. This distinction is even more visible while using the night filter. It's also really interesting to notice how the big holidays follow the same trend of the weekends. The accident rate is less than usual but the severity is higher. For instance, we easily

distinguish the 4th of July (thanksgiving), the 23rd, 24th and 25th of December (Christmas) and the 31st of October (Halloween).

Finally, in the bottom-right corner, the parallel plot helps us distinguish the characteristics of the accident between the different states. At first glance, this plot gives the impression that the states don't differ a lot from each other, but follow common trends. Indeed, apart from some exceptions for each characteristic, the lines are all close to each other.

"Understand if there is a correlation between the weather and environment conditions (presence of clouds, rain, speed of the wind, humidity, etc.) and the car crash characteristics"

To answer this question we can move to the third dashboard (Figure 3). Looking at the correlation bar plots on the overall data, we can notice a couple of things: The wind chill seems to be the most influential among the characteristics while considering queue length and duration. On the other hand, the period of the day (day/night) seems to be positively correlated to the severity. Other relevant results cold be the temperature on the distance and the wind speed on the duration. Nevertheless, the low values of the Pearson correlations (all under 0.10) suggest a difficult linear interpretation of the environmental characteristics in a data mining scenario. They suggest the use of a polynomial machine learning model in case of a car crash's characteristic prediction objective. On the other hand, if we filter for the different states, the scenario changes and reveal strong correlations. These could be exploited and used a state-specific prediction tool.

Analyzing the line plot for weather conditions, we instantly notice how the cloudy type of environment is the most correlated to the number of accidents, followed by the clear, rainy, foggy and snowy type of environment. Overall, all the type of settings seems to cause a constant number of accidents over time. The only deviations from this trend are noticeable in the increase in the middle period of 2019 of the clear environments accidents, which managed to take the place of the cloudy ones, and in the seasonability of the snowy one.

"Understand which are the topology of the most common roads related to car crashes and how these topologies are related to one another."

To answer this question we can move to the fourth dashboard (Figure 4). Here the user can find all the material needed. At first sight one thing should immediately pop up into the user mind: the road setting related to the red color is clearly the most common ones, which are traffic lights. As we can see from the plots, the traffic light is the road setting most wide-spread across cities and indeed the most common road setting affected by car crashes from 2017 to 2019. Right after traffic lights, looking at the time trend, we can see that the color blue and light blue are the second and the third most affected road setting among those years. The light blue color stands for crossings, while the blue color stands for junctions. From this analysis, we could imagine that the location where the accidents are more likely to happen is where two roads meet and where the traffic is regulated by a traffic light. All other minor road settings don't seem to weigh that much on the number of total accidents. Another important reflection that we can develop looking at the dashboard is a comparison between cities. The user can say that Huston, Charlotte, and Raleigh are the cities most affect by accidents close to particular road settings. How can this be possible? We didn't answer that question but we let the user be aware of this information. Another important thing is that, while all cities have the same trend of accidents occurred nearby an amenity location, Austin and Seattle deviate significantly from the cities mean. It could be interesting to understand why this is happening. Do Seattle and Austin have a bad road structure? Do they have amenity zones too close to dangerous roads? The analysis of these aspects can really help cities like Seattle and Austin to decrease this parameter. Last but not least, on the top right of the dashboard, we can find the correlation coefficient that road settings have with each other. From those plots, we can end up with another reflection. Traffic lights, crossing, and amenity zones are more likely to occur together, whilst junction seems to be more independent and unrelated to those settings. This could be because road junctions are likely to be developed outside residential zones, while all the other road settings are a common road structure of cities and small towns. From what we ended up with previous dashboards, we can also assume that junctions, as they are likely to be located far away from the city center, are more likely to have a high severity rate if compared to all other settings.

"Understand how these accidents are handled by the authorities, how much it usually takes to restore the normal traffic flow and which are the geographical areas that struggle the most."

Accidents can not only kill or injure people but also produce traffic congestion problems, and the authorities must be efficient in restoring the normal traffic. If we look at the fifth dashboard (Figure 5), the user will be free to examine what differences are there in handling and restoring the normal traffic (captured by the value "duration" and "distance") across cities and compare them with the country trend. It is also possible to understand if a high duration rate is always related to a high distance rate in a certain zone. What we can understand from the dashboard is that handling car accidents during the night is, as we could imagine, more difficult, but that is not the most important finding. What we can observe from the duration-related map is that at the top left part of the country, with a focus on the state of Oregon, accidents seem to last longer (twice the country trend), and therefore authorities seem struggling with restoring the normal traffic and give support to the injured. A suggestion could be to investigate why this is happening and help, if necessary, the local authorities. Selecting on those cities and looking at the time-related chart we find that this problem seems to affect the state of Oregon from August 2018. Further research can explain why this could have happened. On the other hand, distance seems to be more stable across both cities and time slots, with the country trend slightly increasing over time. An important thing that could pop up into the user mind is that looking at the map, the state of Minnesota and Iowa are more affected by long queues even if authorities are efficient in restoring the normal traffic. The time-related chart, located under the map, suggest a pick of long queues on those states during 2018. The state of Oregon, instead, doesn't seem to suffer from long lines, even with a bad traffic restore. These are all information that the US authorities can use to help each other and to oversee the country trend.

5 Discussion

Since we chose a very big dataset (1GB), we decided not to implement the visual framework to work on, and consequently, we focused on the analysis of the data rather than on the implementation of the whole framework. We chose Tableau [2], as a visual framework, since it is one of the main interactive data visualization software company worldwide, and let the analyst come up with innovative plots. Moreover, we could dig into many geographical visualizations that would have been rough otherwise to implement.

As we said, we focused part of our analysis on map exploitation and map interaction within the same dashboard. We think that maps and geographical data have been exploited properly, as the user can immediately understand the overall trend of some accidents characteristics across the US, and link them to particular cities. Comparison between cities, states and different time slots have been also one of the main activity of our analysis. 90% of our plots can be filtered by filters or by interacting (hovering or selection) with the plot itself. The overall trend can hide some useful information that can be found with a more detailed analysis. The user can analyze trends between cities, states and decide what the next action would be. In the introduction of this paper, we said that the final user of our framework could be used by public administration, the government of the USA, or by the police. This framework raises awareness about those places in the USA which struggle with car accident management and have to be helped by the government. Moreover, the framework not only shows the data itself but analyzes also some possible correlation between car accidents, road settings, and weather conditions. In this way, public administration will be able to understand when and where most of the accidents are likely to occur, and how to design safer roads. This feature can be found in

the third and fourth dashboard which present all correlation coefficient visualized in colored bar charts. We have analyzed the dataset as better we could, exploiting most of the fields to answer the questions we named previously. Not all the possible questions have been answered, but we tried to capture the most interesting ones. One negative thing about the analysis is that the dataset is really big and could not let everyone enjoy the analysis. If the user doesn't have enough computational power he might be annoyed about the user experience, as the framework might slow down and require some time to elaborate the enormous amount of data. Moreover, sometimes trends "jumps" over time. This means that the data sometimes has fast changes over time and deviates significantly from its previous trend. We were not able to say if these strange jumps could or could not have been outliers. It can be that a company has extended significantly the input sensors coverage or that the trend actually increased that much. We recommend further analyses to better explain these jumps.

6 Conclusion

In the USA, car accidents happen in huge amounts every day. The ones that have to deal mostly with car crashes and their consequences are local authorities and the police. With this paper, we wanted to present a new framework that could help them to have an overview of the problem and a clear understanding of what could influence the appearance of car accidents. Are they likely to happen in a certain zone? Which zones are most affected by the consequences that a car accident can bring? We tried to answer all of these questions and gave the authorities a framework that could help them to act where is most needed. Within the framework, the user can look up, compare, and find out brand new information that wouldn't have been possible by looking at the giant amount of data itself. We think that such a framework could really improve the efficiency of the authorities and can find out many management problems of the local ones across the country. Some of the main problems we found while conducting our analysis can be summarized as follow:

- Overall the trend of the number of accidents occurred per year is growing. California, Texas, and Florida are the States most affected by injuries.
- Car accidents take place most of the time during weekdays, meanwhile, they decrease in number during the weekends and national holidays. However, these lasts, have a greater impact on the traffic flow.
- Car accidents that occur in the highways are usually far more severe than the one happening in other types of streets. This is probably due to the velocity with which cars are used to travel in these types of streets, and consequently in the more dangerous scenarios that car crashes produce.
- The cloudy type of weather environment is the most correlated to the number of accidents, followed by the clear, rainy, foggy and snowy type of environment.
- The road where the accidents are more likely to happen is where two roads meet or where the traffic is regulated by a traffic light. All other road settings don't seem to influence the number of total accidents.
- The cities of Huston, Charlotte and Raleigh are the cities most affected by accidents occurred close to particular road settings. On the other hand, Seattle and Austin have high amounts of accidents that happen close to amenity zones.
- Traffic light and crossing are positive correlated. Junctions seem not to be correlated to other type of road settings.
- The state of Oregon seems that needs help on car accident management, as the time to restore the traffic to the normal flow is way higher if compared to the country average.

Work division

In order to have equal roles, we divided the development of the dashboards as follows: Leonardo Remondini developed dashboard number 1, 4 and 6; Mattia Molon developed the other dashboard, that means dashboard number 2, 3 and 5. A similar equal division has been used while writing the paper. Overall, we believe to have divided the work in a balanced way between the two members of the team.

References

[1] S. Moosavi, M. H. Samavatian, S. Parthasarathy, and R. Ramnath. A countrywide traffic accident dataset. *CoRR*, abs/1906.05409, 2019.

[2] T. software. https://www.tableau.com/, January 2020.

[3] F. . Talwar. Car accidents statistic 2013. https://www.fishertalwar.com/car-accident-statistics, December 2013.