Analysis and Visualization of Road Traffic Accidents in USA

Vineeth Reddy Chinthala
Department of Applied Data Science
San Jose State University
San Jose, USA
015294591

***Abstract*— In the recent years, there has been a significant increase in the number of road accidents, with a yearly average death rate of 1 million individuals. While several factors play a role in car accidents such as driving under influence, speeding, distracted driving etc., traffic is one of the most consequential factors. This project provides a visualization of car accidents caused in USA due to road traffic after thorough analysis and assists in decision-making for drivers and police alike.**

Keywords—road accidents, data analysis, information visualization, dashboard, storytelling

I. Introduction

Recently, the World Health Organization (WHO), published a factsheet detailing the consequences of road traffic incidents and its severity. Road traffic accidents affect the Gross Domestic Product (GDP) of a country, contributing to a three percent loss on a yearly basis [1]. Accidents caused due to road traffic not only negatively influences the driver's life but also other defenceless road users such as bicyclists and pedestrians. The United Nations has made a commitment to halve the number of deaths caused due to road traffic by the year 2030 [1].

In USA alone, more than 38,000 people die every year due to road accidents. For every 100,000 residents, the U.S. traffic fatality rate is 12.4 deaths [2]. According to the road crash statistics provided by the Association for Safe International Road Travel (ASIRT), U.S. suffers the highest death rate, nearly 50% higher, as compared to its contemporaries.

A. Motivation

Several reasons contribute to road crashes, such as, reckless driver behaviour, improper road infrastructure, lack of enforcement of traffic rules, and insufficient post trauma care [2]. With ownership of car becoming a more affordable occurrence, the rate of accidents due to road traffic has also proportionately increased. With ownership of car becoming a more affordable occurrence, the rate of accidents due to road traffic has also proportionately increased. While road traffic crashes cannot be completely eradicated, their frequency can be reduced through conscious decision making. To assist with this, data analysis and visualization come into picture.

The prime target of accident analysis is prevention. Discovering the reasons for an accident and finding a way ways to control or eliminate of it can assist with keeping comparable accidents from occurring later on. When the causes are set up, precautions should be distinguished and executed to prevent a recurrence.

B. Proposed Method:

Daily, accidents are recorded into a database which is made publicly available but most of it is not understandable to a common man. Therefore, in this project, data is sourced from public databases and using Tableau, a dashboard comprising of time-specific incidents, locations and days of the week shall be projected for the end-user to view. Being an interactive dashboard, the end-user or targeted audience, in this case, regular drivers and police, can make calculated decisions based on the visualizations resulting in fewer accidents and more cautious driving.

II. Related Work

Visual Analytics for road traffic accidents came into picture in the mid 2010s. In 2014, Soltani and Askari [3] used Geographic Information Systems (GIS)-based spatiotemporal visualization techniques to analyse intra-urban traffic accident patterns. The analysis aimed to identify accident locations with a high accident rate and poor safety areas using the Kernel Estimation Density (KED) method. It was observed that main roads are essentially prone to a higher rate of accidents and the most sensitive hours for road crashes were during peak traffic congestion. Accident-prone locations are mainly located in neighbourhoods with a higher speed and traffic. Therefore, they should be considered as the priority research sites for security promotion programs.

A second comparative study conducted in 2014 [4], also took up the task of analyzing large scale vehicle traffic data to forecast the reasons behind traffic related accidents using information visualization techniques. Various information visualization techniques were discussed, such as, Geometric Projection, Pixel-oriented, and Hierarchical visualization. It was concluded that while the integration of Information visualization with Traffic Engineering had seen a significant growth, there is still scope for improvement in terms of visual overlapping and integration of different traffic networks.

In 2015, another study [5] was conducted to identify the black spots, which are otherwise known as road accident accumulation zones and the location of the study was Portugal. Visual analytics techniques were used to recognize the displacement of accumulation zones on annual fixed sliding windows. Three main parameters are taken into consideration into this study, namely, number of minor injuries, number of severe injuries, and number of fatalities. The following equation was used to calculate the severity indicator [5]:

$$
\begin{aligned}
&severityindicator \\
&= 3 \times number\ of\ minor\ injuries \\
&+ 10 \times number\ of\ severe\ injuries \\
&+ 100 \times number\ of\ fatalities. \qquad (1)
\end{aligned}
$$

The criteria for determining a road accident accumulation zone were decided as any zone where the value of the severity indicator was greater than 20.

Another recent study conducted in the year 2020 by researchers [6] concentrated on the Hamilton County, specifically the Chattanooga Metropolitan Area located along the Tennessee-Georgia border. The researchers used spatial visualization techniques to determine safety patterns in various traffic environments [6]. For better understanding and comparison of spatiotemporal figures, a textured tile calendar was built. Heat-maps for traffic incidents recorded by 911 were charted for a period of one year and it was analyzed that other factors such as weather and natural disasters also have an impact on traffic.

III. Method

A. Dataset

Several steps assist in building this interactive dashboard, but the first and foremost, is sourcing the dataset and ensuring that the data is cleaned prior to loading it into Tableau.

The dataset used in this project has been sourced from Kaggle [7]. The dataset comprises of 1.5 million records with 47 attributes. It is a country wide road accident dataset and covers nearly 49 states within the United States. The data has been collected using multiple APIs that dispense streaming traffic data. The columns of interest for this project are, namely, severity of the accident, the start times and end times, states and cities, days of the week, and major points or accident-prone zones.

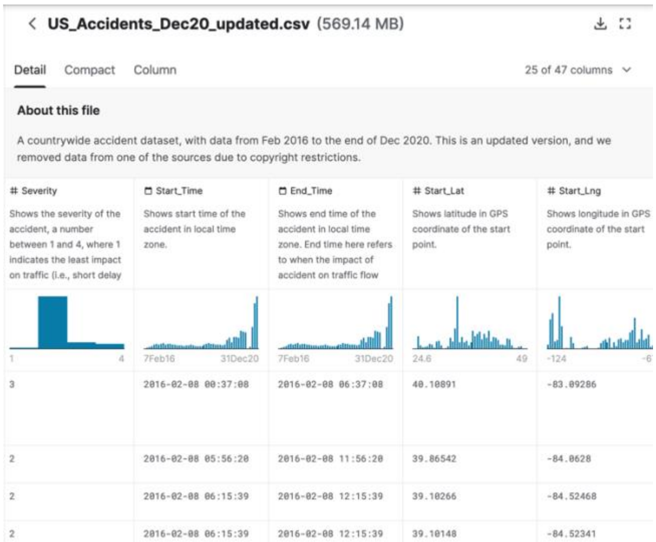

Fig. 1: A snapshot of the dataset that has been sourced from Kaggle

B. Data Pre-processing

To ensure there is no data overlapping, null values or redundancies, the dataset was cleaned using Python pandas. On processing the data, it was found that for the year 2020, incorrect values had been observed and were not in sync with actual factual data that can be found on the internet through other resources. After these incorrect values were filtered out, the number of records reduced by nearly 700,000 leaving us with 730,000 records. The updated dataset has records for the years between 2016 and 2019. Any out-of-range values and null values were removed first, after which the dataset was checked for any redundancies. Once the redundant data was removed, the dataset was reconverted and saved as a CSV file for future data manipulation, analysis, and visualization.

C. Tableau

Once the data cleaning process was completed, the updated CSV file was imported into Tableau. Due to the dataset consisting of nearly 730,000 records, it took around four to five minutes for the import command to be executed. Once the file was imported, the necessary data analysis was performed.

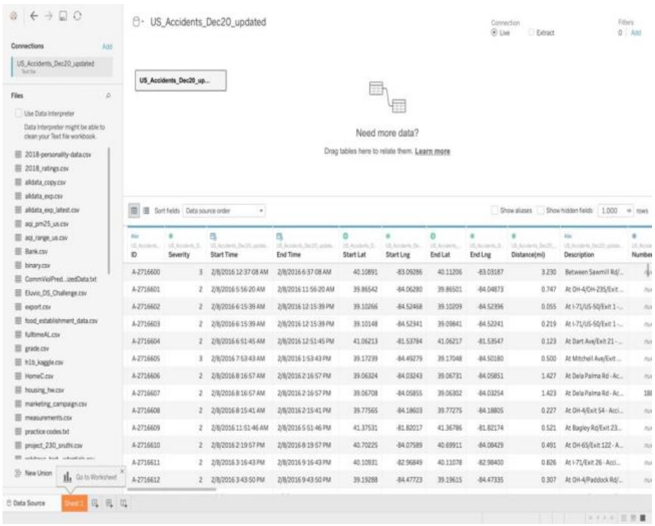

Fig. 2: Updated dataset has been imported to Tableau

D. Data Analysis and Visualization

Sets, bins, and calculation fields were used to develop the required charts for the dashboard. While charts must be visually appealing to gain the attention of the end-user, it is equally necessary to be able to communicate the meaning of each chart in an uncomplicated fashion. To ensure this clarity is provided throughout the dashboard, the following has been performed:

- Case conditions were applied to check whether a particular day would be categorised as a weekday or as a weekend.
- Manual sorting was performed to sort weekdays and weekends. As can be observed in the upcoming charts, the week begins on a Monday and ends on a Sunday.
- For temperature, bins with a width of ten were created to show results.

- Using the 'Dashboard Action' feature for each chart and adding the necessary filter, to make the dashboard interactive.

Moving forward, various charts and their description has been provided for better understanding.

IV. Results

A. *Severity-level based Accident Analysis*

Figure 3 depicts the number of accidents based on their severity level. The severity level ranges from 1, which is the lowest to 4, which is the highest. Due to the discrepancies in the records for the year 2020, the updated dataset has values recorded for only the years between 2016 and 2019. Hence, we see that in this pie chart, there are only three divisions ranging from levels 2 to 4, of severity.

A donut chart was used in this scenario, since proper bifurcations can be projected. A donut chart also emphasizes on the proper use of space while also providing meaning to the chart. The 'Select' feature was used to display the content of each division.

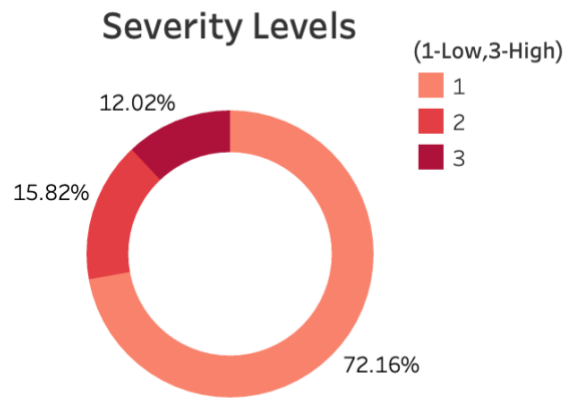

Fig. 3: No. of accidents based on the severity levels between 2016 and 2019

B. *Accidents based on Type of Day (Weekday/ Weekend)*

Figure 4 shows the calculated field that was used to determine whether a given day was a weekend or a weekday. If the start time ranges between 2 and 6, then the day is considered to be a weekday, else, it is considered as a weekend.

Fig. 4: Condition to determine weekday or weekend

Figure 5 shows the accidents by hour (weekday/ weekend). Each bin has a width of two hours. A line graph has been chosen in order to initiate a continuous comparison between the pattern in which accidents occur during both weekdays and weekends.

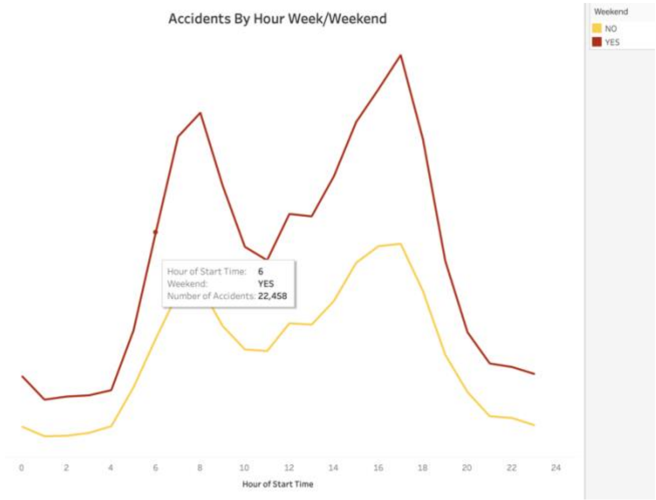

Fig. 5: Accidents by hour (weekday/ weekend)

C. *Accidents based on Timezone*

Figure 6 shows the accidents based on time zone and the respective severity levels (weekday/ weekend). There are mainly four main time-zones that have been taken into consideration here, the Pacific Standard Time (PST), the Eastern Standard Time (EST), the Central Standard Time (CST), and the Mountain Time (MT). A bar graph was used with a colour shading to depict different severity level, with the darkest one being 'critical'.

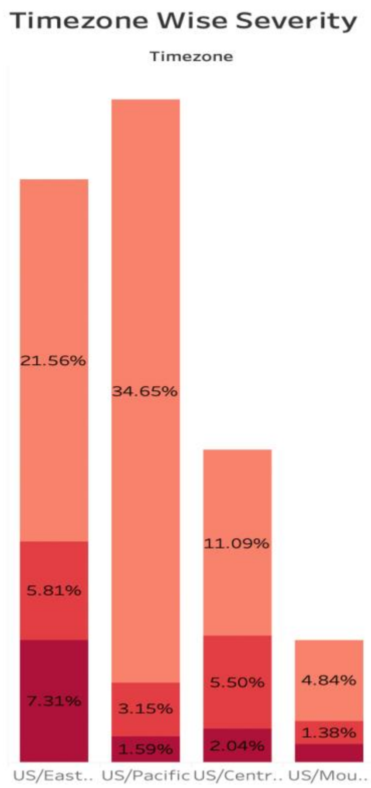

Fig. 6: Accidents based on the time zone and its severity level

D. *Zones accidents are likely to occur*

Figure 7 shows the calculated field to determine the potential zones which are highly accident prone. The bar graph is only created when the following if/else condition is fulfilled else the value of ‘None’ will be printed and it is better to avoid null values for better data analysis and visualization.

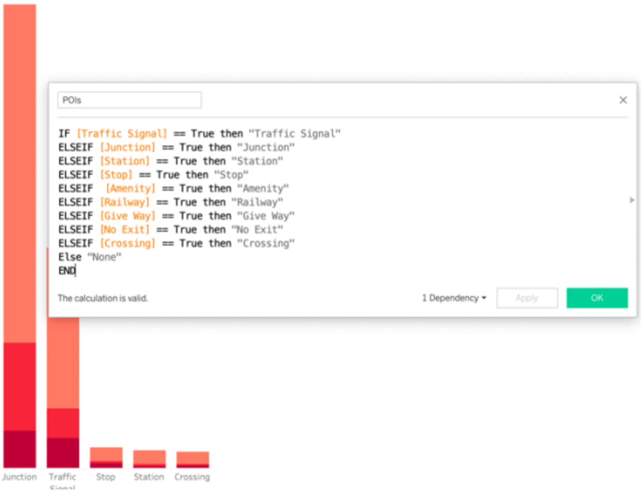

Fig. 7: Filter to determine major accident-prone zones

Figure 8 shows the major accident-prone zones. It can be seen that mostly 'junctions', followed by 'traffic signals' are prone to accidents. The various colors in the bar graph are to differentiate between the severity levels of the accidents that occur. As the color ‘red’ represents danger, the darkest shade of ‘red’ is the most critical level of accident. Here, only the top five zones have been considered between the year 2016 and 2019.

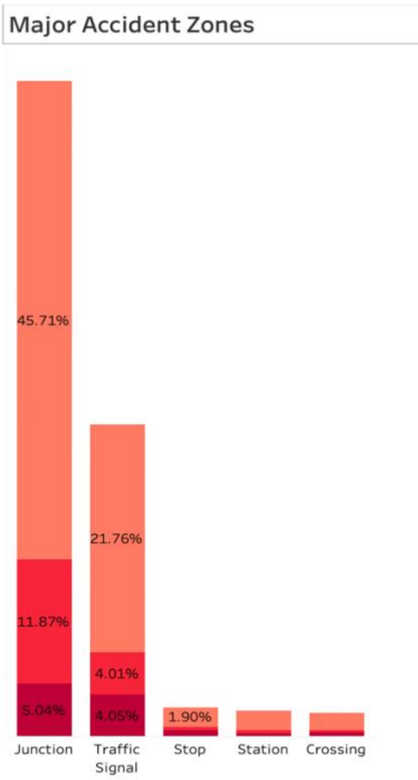

Fig. 8: Major Accident Zones

E. *Top states with the highest number of accidents*

Figure 9 and 10 shows the filter used to view the top ten states in the country that has the highest number of accidents. It can be observed that the filter is names 'IN/OUT', this is to show which states are present within the top ten and which are not.

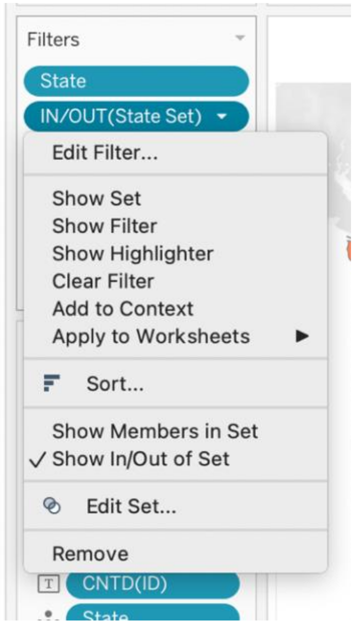

Fig. 9: Filter to determine major accident-prone zones

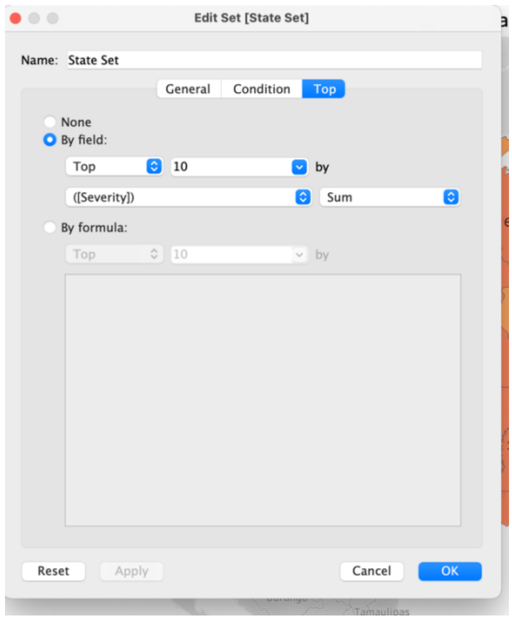

Fig. 10: Condition to only highlight the top 10 states

Figure 11 shows the map graph highlighting the top ten states in the country in a 'light orange' shade and the remaining states in a 'dark orange' shade. This was done to ensure that a comparison can be drawn between all the 49 states for which the data is available between the year 2016 and 2019. A map graph was chosen here for the purpose of clarity and visual appeal.

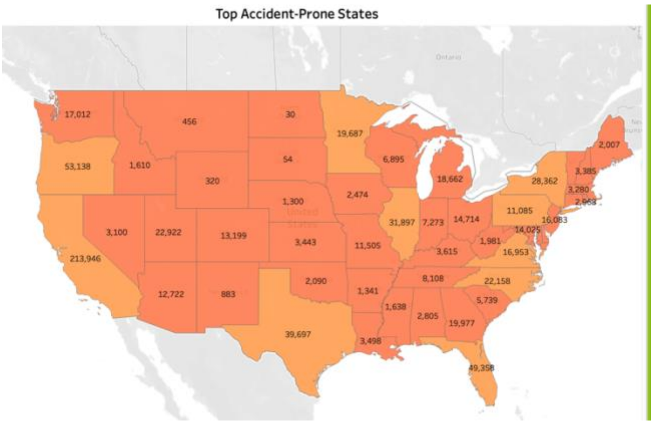

Fig. 11: Top 10 states with highest number of accidents

F. Top cities and highways based on highest number of accidents

Figure 12 shows the top 30 cities with the highest number of accidents and their severity level to increase the scope of analysis and to ensure no detail is overlooked. A horizontal bar graph has been shown here, with the x-axis representing the number of accidents and the y-axis representing the cities. The severity levels can be understood through the varying color.

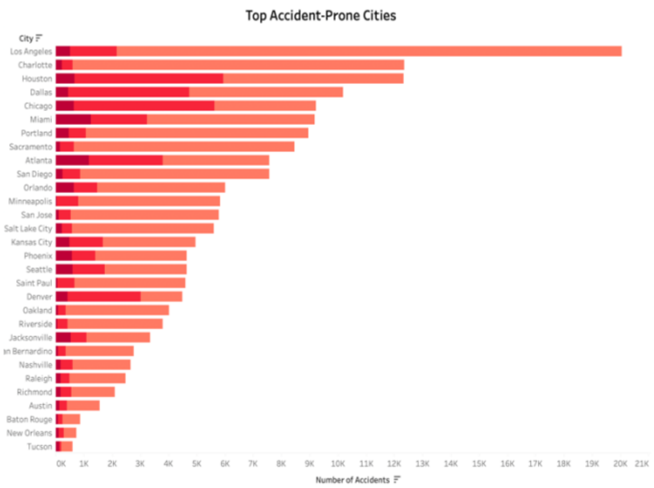

Fig. 12: Top 30 accident prone cities and their severity levels

A similar concept and intention has been applied to obtain a graph for the top ten accident-prone highways.

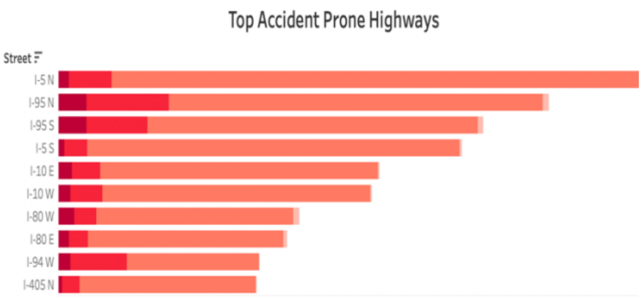

Fig. 13: Top 10 accident prone cities and their severity levels

G. Accidents based on temperature

Figure 14 shows the number of accidents that occur due to changes in temperature. For clarity, the temperature has been divided into bins with a width of 10 degrees Fahrenheit. The x-axis represents the temperature in Fahrenheit and the y-axis represents the number of accidents.

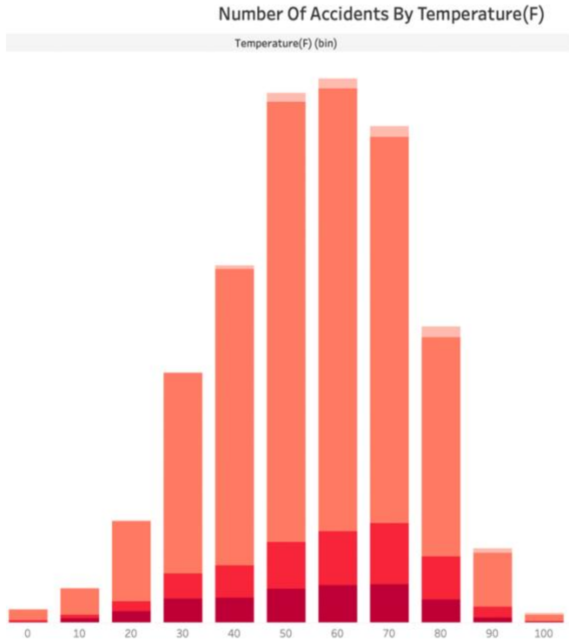

Fig. 14: Top 10 states with highest number of accidents

H. Accidents based on the weather condition

Weather conditions play a crucial part in determining the critical nature or severity of an accident. In this project, the top ten weather conditions have been considered, these are namely, fair weather, cloudy (high, normal, partly, and scattered), clear weather, rain (light and heavy showers), snow, and fog. The x-axis represents the weather condition, and the y-axis represents the number of accidents.

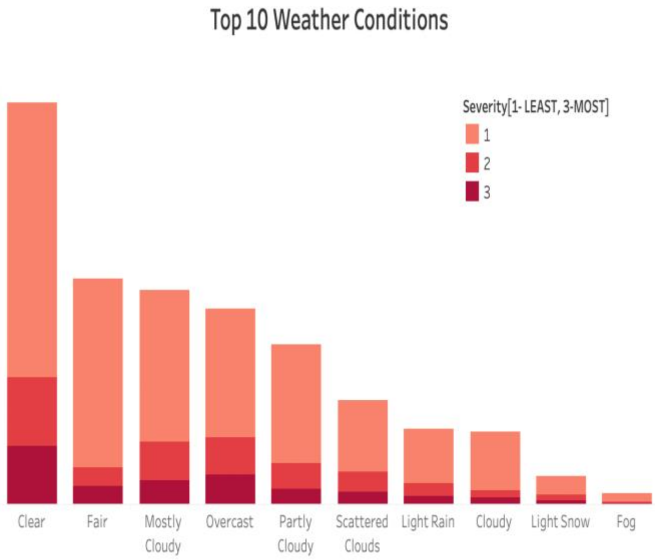

Fig. 15: Top 10 weather conditions and its effect on accidents

I. *Dashboard*

An interactive dashboard has been created by combining several worksheets, using the 'Dashboard Action' feature for each chart and adding the necessary filter, to make the dashboard interactive.

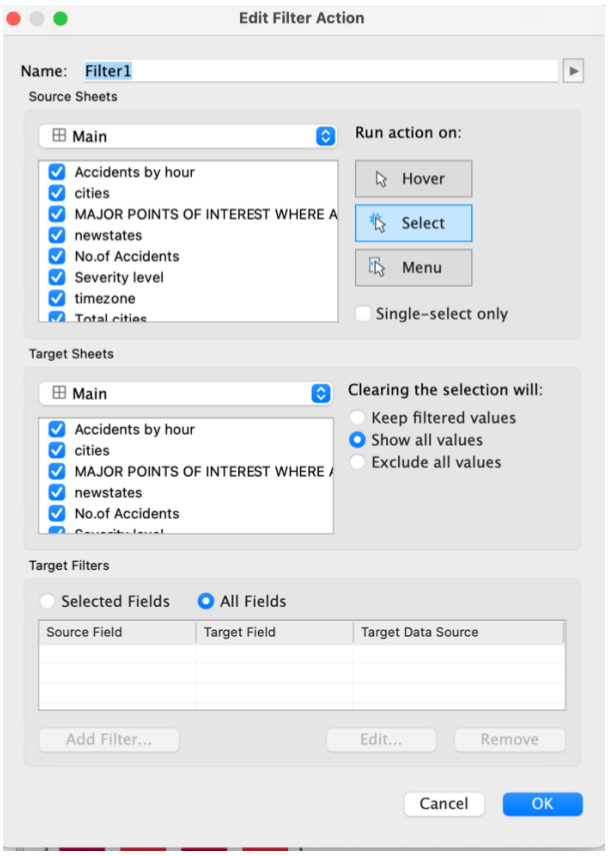

Fig. 16: Dashboard filter action for interactive dashboard

The interactive dashboard has been enabled with the 'Select' feature, therefore on clicking any of the components in the graphs, the charts in the dashboard changes in accordance. The dashboard displays, the total number of accidents, the total number of states, and the total number of cities as reference. It also contains various bar graphs, map graph, line graph, a pie chart, and a donut chart. Additionally, the number of accidents per year has been depicted using a pie chart.

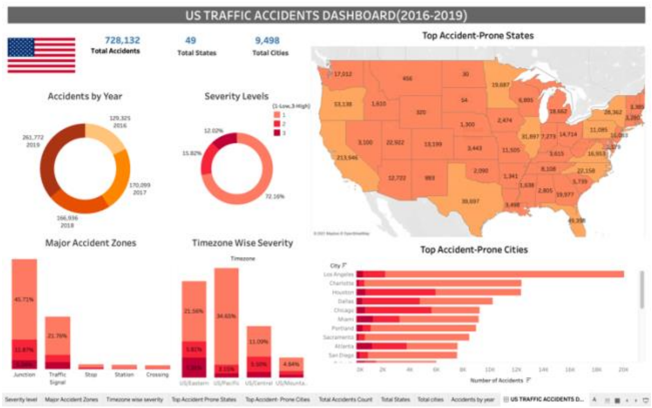

Fig. 17: Interactive Dashboard View

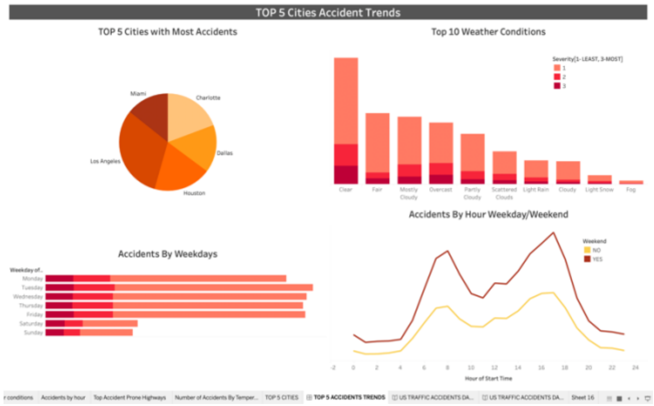

Fig. 18: Interactive Dashboard View – Top 5 cities and their accident statistics

V. DISCUSSIONS

From the donut chart in **Figure 3**, it can be deduced that 75% of the accidents that occur within severity level 2.

Through the chart in **Figure 5**, maximum number of accidents occur during the early morning hours or late evening hours in the weekends, since this is the time when people leave for parties, clubs, and restaurants or return from these places. On weekdays, accidents are more likely to occur during commuting to and from the office especially through peak traffic hours, between 7.30 am to 8.30 am and between 5.00 pm to 6.00 pm respectively.

In **Figure 6**, it can be observed that majority, nearly 34.65% takes place in the Pacific time zone and is of severity level 2. This can be seen due to the lack of proper public transport infrastructure as opposed to the east coast, therefore, requiring more and more individuals to own a car for daily commute.

In **Figure 8**, mostly 'junctions', followed by 'traffic signals' are prone to accidents. This is due to distracted driving, when individuals do not pay attention to all sides of the roads as well as pedestrians. Such irresponsible behavior proves to be dangerous to pedestrians and other drivers as well.

From **Figure 11**, it can be clearly seen that California holds the record for highest number of accidents. The state of California has the highest traffic accidents (213,946) with almost 4 times the records compared to Oregon, the second highest accident-prone state (53,138). This is because there are several highways that are highly packed and rash driving can be seen on such highways due to the increasing frustration in waiting for long periods of time in the traffic.

From the bar graph in **Figure 12**, Los Angeles has the highest number of accidents, and this is caused due to the several car chases as well as rash driving that takes place on the I5 North highway.

Figure 13 shows that I5 North highway is highly prone to accidents. As mentioned previously, this highway connects San Diego to Los Angeles and extends further north up to San Jose. Since this is the quickest connection between Los Angeles and San Jose, drivers tend to choose this route and opposed to Highway 1 by the Pacific Coast. Due to heavy traffic and intense tailgating, the chances of accidents on this route are quite high.

Figure 14, one can observe that the number of accidents is inversely proportional to the temperature, i.e., as the temperature decreases the chances of an accident increases.

This is due to the possibility of fog formation at lower temperatures that could decrease the visibility of the driver.

In **Figure 15,** it is observed that higher number of accidents take place during fair weather. Fair weather stands for a normal, sunny day and this drives more families and individuals to venture out to make most of the weather. Travel is more restricted during other weather conditions as is obvious.

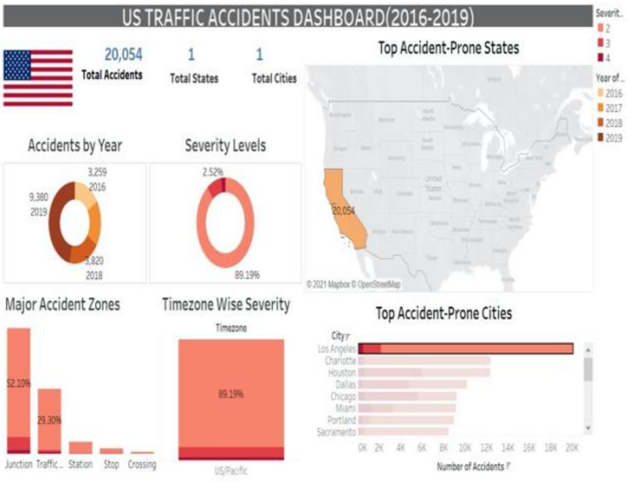

Fig. 19: Interactive Dashboard View – Los Angeles

Los Angeles is known for its notorious car racing what with the prevalence of several gangs. While car chasing might be adventurous for some, it is not for majority of the other commuters.

In **Figure 19**, the dashboard has been generated specifically for Los Angeles. As was seen, California has the highest accident numbers. For a fact, any weekday of December between 7am to 10am and 2pm to 7pm is considered to be the most dangerous times to drive, especially in the state of California. The dashboard also displays the number of accidents that have taken place for each year between 2016 and 2019.

In the year 2019, one such incident left the regular riders of I-5 disturbed. I-5 is one of the most busiest highways in USA with multiple lanes. Due to heavy traffic and the urgency to reach a destination on time, several commuters constantly change lanes. On one such day, during heavy traffic, an irresponsible rider decided to change his lane, without ensuring if there was safe distance from cars in the left lane. While switching lanes, the car crashed into another vehicle causing severe damage to both the car and the driver. Due to the sudden brake and force, this caused a multi-car accident and it was extremely difficult for the police and ambulance to reach the spot proving dangerous to the passengers involved in this horrendous accident.

Many such accidents which are extremely severe in nature occur till date. So many factors, such as distracted driving, rashness and impulsive nature of riders have left many commuters extremely scared to venture out on highways such as I-5. This fear needs to be acknowledged and necessary steps need to be taken. This is where our dashboard comes into picture.

A. *Insights and Preventive Measures*

Some of the insights we gained through the visual analysis of road traffic accidents are as follows:

- Measures need to be taken to reduce or divert traffic, especially during peak hours and on weekdays.
- Alternate and safer routes must be made accessible for drivers to avoid accident-prone highways.
- Rules must be more strictly enforced when it comes to distracted and rash driving.
- Proper road infrastructure must be built for safe travel especially in accident-prone areas.
- Roadside cameras must be present at every couple of miles ensuring discipline amongst drivers.
- Warnings or alerts about the weather conditions must be sent to drivers and police alike to provide sufficient time to make alternate arrangements.

VI. Future Work

The dataset for the year 2020, had multiple discrepancies in the data. It was observed that there was a sudden drop of accidents in the month of July and August, whereas a sudden rise in September and the following months. While this could potentially be due to festive season and slight relaxation in rules in the latter part of 2020. Therefore, sourcing the correct data and implementing the dashboard in real-time is of priority.

Also, converting this interactive dashboard to a mobile application, for easy access to our targeted demography of drivers, pedestrians and police after adding a few more features, such as an integrated weather application with the dashboard.

What with the current youth of the world more active on mobile and social media platforms, such an application would be very useful, while also sending out notifications about the traffic and tracking travel habits to advice the best route.

VII. Charting

Through the lessons taught in the class, we were able to better understand and implement the following concepts:

- Ensuring that the graphs are visually pleasing and are self-explanatory, i.e., each graph holds a meaning in itself, thereby corroborating that expressiveness and effectiveness is intact.
- Each graph has been placed in an unique position to make sure there is no graphical overlaps.
- A colour-contrast test has been performed on the dashboard and it returned a 98\% accessibility.
- **Proximity:** The dashboard has been divided into different subsets having similar properties within the subsets.
- **Similarity:** Graphs with similar visualization have been grouped together. For instance, vertical bar graphs are present together.
- **Continuity:** In the figure 5, it can be observed that there is a comparison between line graphs and they are in alignment

VIII. NOVELTY

From the research that has been conducted for the literature survey, a simple interactive dashboard has never been created. While various Geographic Information Systems (GIS) based systems are prevalent and visual analytics has been performed, but they have never been connected and minute details such as the effect of weather and temperature on accidents haven't been accounted for, thus making our interactive dashboard a first of its kind. Through the help of this interactive dashboard, regular drivers and pedestrians alike can understand the following:

- Accident prone areas that require extra caution while travelling through, by both, drivers, and pedestrians.
- Areas that require close police monitoring and an ambulance on standby.
- Awareness on the weather conditions and its severity.

ACKNOWLEDGMENT

We would like to thank our Data Visualization professor, Dr. Abbas Moallem, for giving us the opportunity to learn and implement the concepts learnt in class using Tableau and for providing us guidance throughout the project. Would also like to acknowledge our TA, Ms. Swathi Badicole, for clarifying any questions we had regarding the project report and other logistics.

REFERENCES

[1] "Road traffic injuries," World Health Organization, 21-Jun-2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] "Road safety facts," Association for Safe International Road Travel, 25-Aug-2021. [Online]. Available: https://www.asirt.org/safe-travel/road-safety-facts/.

[3] A. Soltani and S. Askari, "Analysis of intra-urban traffic accidents using spatiotemporal visualization techniques," Transport and Telecommunication Journal, vol. 15, no. 3, pp. 227–232, 2014.

[4] H. W. A. S. Gondim, H. A. D. do Nascimento and D. Reilly, "Visualizing large scale vehicle traffic network data a survey of the state-of-the-art," 2014 International Conference on Information Visualization Theory and Applications (IVAPP), 2014, pp. 337-346.

[5] L. Ramos, L. Silva, M. Y. Santos, and J. M. Pires, "Detection of road accident accumulation zones with a visual analytics approach," Procedia Computer Science, vol. 64, pp. 969–976, Dec. 2015.

[6] A. S. Berres, H. Xu, S. A. Tennille, J. Severino, S. Ravulaparthy, and J. Sanyal, "Explorative visualization for traffic safety using adaptive study areas," Transportation Research Record: Journal of the Transportation Research Board, vol. 2675, no. 6, pp. 51–69, Jan. 2021.

[7] S. Moosavi, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and R. Ramnath, "Accident risk prediction based on heterogeneous sparse data," Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Sep. 2019.

[8] M. Ehline, "Los Angeles car accident statistics: Safety, Tips and data discussed," Ehline Law Firm Personal Injury Attorneys, APLC, 08-Jul-2021. [Online]. Available: https://ehlinelaw.com/practice/car-accident/traffic-accident-faq/los-angeles-and-u-s-statistics.