Shashank Reddy Gollapally
INFO 5709
11376637

Analysis of US Accidents

Introduction:

The lives of many individuals and societies have been enhanced through motorization. But the benefits always come with a price. Over 1.35 million people die in road crashes approximately, each year, an average of 3700 people lose their lives on the roads, globally. This road crashes often result in long-term disabilities as an additional 20-50 million people suffer non-fatal injuries. One of the leading cause of deaths in the U.S. for the people aged 1-54 is the road crashes. It costs U.S. citizens $871 billion due to the economic and societal impact of road crashes. Road crashes act as the single greatest annual cause of death of the U.S. citizens traveling abroad. The purpose of this project is to analyze the accident data to discover something useful, looking at the severity of road accidents.

Related Work:

A numerous application such as studying accident hotspot locations, real-time accident prediction, studying the impact of precipitation or other environmental stimuli on the accident occurrence or causality analysis and extracting cause and effect rules to predict accidents can be applied to resolve the issue on U.S accidents. A great deal of research is going on out there that addresses accidents all over the world. By far, most of them are on accident analysis, and the prediction has used restricted assets that do not give a full idea of the issue and affecting the results that we are seeking. One of the research papers (A Countrywide Traffic Accident dataset) tried to address this issue by collecting the information from API resources through several sources. Another paper by Eisenberg carried out the analysis in identifying the impact of road accidents with a large dataset from 1975 to 2000 of about 456000 crashes in 48 states of U.S. To analyze real-time traffic accident prediction, recent studies by Najjar et al. (Najjar, Kaneko, & Miyanaga, 2017) used large scale datasets.

Apart from these studies a paper published by Jie Wang, Yozhong Shen, Asad Khattak, has done visualization in Geographic Information Systems (GIS) that greatly facilitate handling in many aspects. The world leader in GIS research and development is the Environmental Systems Research Institute (ESRI), Inc. A software package developed by ESRI, ArcGIS has been used to display events associated with a road network that includes accident locations, and pavement quality. Though the paper states the causes of road accidents and their severity, similar to our visualization report, it aims to address the problems that are caused by the ArcGIS. It provides an improved method to utilize the relevant information of road accidents such as route number, direction, and milepost for accurate traffic accident visualization and analysis. For our visualization, we are solely focusing on the factors behind the cause of road accidents.

Data Description:

There is a dataset available on the Kaggle website for traffic accidents. The reason behind opting for this dataset is that it is a countrywide traffic wide dataset, which covers 49 states of the country United States. The includes two APIs that provide streaming traffic event data and are continuously being collected from February 2016. Traffic events captured by a variety of entities, such as the transportation departments of US and state, traffic cameras, and traffic sensors, law enforcement agencies within the road networks are broadcast by these APIs. There are about 3.0 million accident records available in this dataset currently.

The analysis includes an ID which is a unique identifier of the accident record, the source of the report, i.e. the API, severity of the accident which is indicated as a number between 1 and 4, 1 being the indication for least impact on traffic such as a short delay resulting in an accident and 4 being the significant impact on traffic such as long delay. One of the other factors that cause road accidents is time, and it has been represented using the start time and end time in the local time zone, latitude, and longitude in the GPS coordinate of the start point and endpoints. Distance, the length of the road extent that is affected by the accident, description of the accident, address field, including number, street, side, city, country, state, zip code, and Timezone based on the location of the accident. The next factor is climatic conditions which include, wind direction, wind speed, precipitation measures in inches, and weather conditions. Factors such as an amenity, speed bump or hump, crossing, give away, junction, no exit, railway, roundabout, station, stops, traffic calming, traffic signals, and turning loops in a nearby location, are also a source of road accidents. Lastly, the period of the day such as sunrise or sunset, civil twilight, nautical twilight, astronomical twilight, are also a cause for road accidents.

Data Cleaning:

The platform used for analysis in Power BI. I have opted python for filtering the dataset as it can be easily understood and a commonly used programming language. It is an object-oriented language used for analyzing data. The reason behind choosing python is that it runs on an interpretive system through which the code can be executed as soon as it is written. The dataset we are using has 3 million records, but it is not yet ready to use for analysis. The dataset has any anomalies such as the date format, null records, duplicate records, and unwanted columns. Therefore, data cleaning is the most important and mandatory step to address all these anomalies in the data. Filtering of data has been in the below analysis,

Date format:

As the data has been collected from many resources, the format of the date was not consistent. It is necessary to have the data in a logical form, therefore all the date formats were changed to "MM/DD/YYYY" for the analysis.

Null Records:

Few null records have been deleted from the dataset. They are replaced with appropriate values considering the percentage of null records depending on the category.

Duplicate Records:

All the duplicate records are deleted to gain more accuracy, as the dataset has many duplicate records.

Unwanted Columns:

All the columns consisting of many null records are deleted to gain more precision from the dataset. Also, the columns that have not been used for the analysis were also removed.

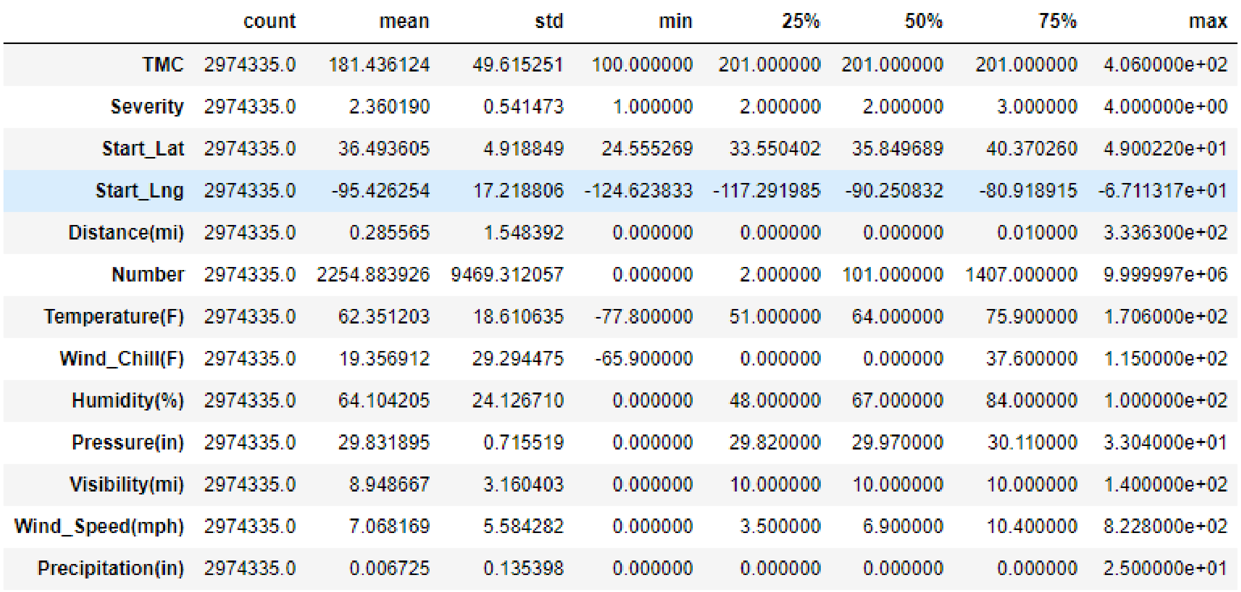
Exploratory Data Analysis (EDA):

Exploratory Data Analysis is an approach that is used for data analysis that employs many techniques such as maximizing the insight into a data set, detecting outliers and anomalies, developing parsimonious tools, testing underlying assumptions, uncovering underlying structure, determining optimal factor settings, and extracting important variables.

Although the term statistical graphics are used with EDA interchangeable, they are not identical. Statistical graphics are a collection of techniques. All the focus is on one characterization aspect based graphically. Contrary to this, EDA ventures into a larger scenario, an approach to data analysis in postponing the usual assumptions about what kind of a model the data follow to reveal its underlying structure and model with the more direct approach of allowing the data itself. EDA is not just a collection of techniques online statistical graphics, it is a philosophy as to how a dataset can be dissected, what, and how we look for, and how we interpret the information. Though EDA uses the collection of techniques referred to as “statistical graphics” heavily, it is not identical to the statistical graphics.

A component of data analytics is statistical analysis. Statistical analysis involves collecting and scrutinizing the data samples in a set of items from which all the samples are drawn in the context of business intelligence. Statistical analysis is used for describing the nature of the data for the analysis, exploring the relation of the data, creating a model by summarizing how the data is related, and understanding it, proving the validity of the model and employing predictive analysis for acting as a guide in future actions. The statistical analysis carried out on the dataset can be viewed in the figure 1 below. We have the following information from the dataset,

The dataset consists of information from 49 states, with a total number of 2974335 traffic accidents. An average of around 0.285 miles (i.e., mean) has been affected by these road accidents. The climatic conditions we have considered for the dataset has a minimum temperature (in Fahrenheit) at -77.8 and a maximum of 170.06 with a humidity reading of 0 to 1. The wind-chill minimum is at -65.9 where the wind-speed starts from 0 and increases up to 822.80. The pressure on the air has been analyzed to have started from a rest state, i.e., 0 and have sped up to 33.04.



	count	mean	std	min	25%	50%	75%	max
TMC	2974335.0	181.436124	49.615251	100.000000	201.000000	201.000000	201.000000	4.060000e+02
Severity	2974335.0	2.360190	0.541473	1.000000	2.000000	2.000000	3.000000	4.000000e+00
Start_Lat	2974335.0	36.493605	4.918849	24.555269	33.550402	35.849689	40.370260	4.900220e+01
Start_Lng	2974335.0	-95.426254	17.218806	-124.623833	-117.291985	-90.250832	-80.918915	-6.711317e+01
Distance(mi)	2974335.0	0.285565	1.548392	0.000000	0.000000	0.000000	0.010000	3.336300e+02
Number	2974335.0	2254.883926	9469.312057	0.000000	2.000000	101.000000	1407.000000	9.999997e+06
Temperature(F)	2974335.0	62.351203	18.610635	-77.800000	51.000000	64.000000	75.900000	1.706000e+02
Wind_Chill(F)	2974335.0	19.356912	29.294475	-65.900000	0.000000	0.000000	37.600000	1.150000e+02
Humidity(%)	2974335.0	64.104205	24.126710	0.000000	48.000000	67.000000	84.000000	1.000000e+02
Pressure(in)	2974335.0	29.831895	0.715519	0.000000	29.820000	29.970000	30.110000	3.304000e+01
Visibility(mi)	2974335.0	8.948667	3.160403	0.000000	10.000000	10.000000	10.000000	1.400000e+02
Wind_Speed(mph)	2974335.0	7.068169	5.584282	0.000000	3.500000	6.900000	10.400000	8.228000e+02
Precipitation(in)	2974335.0	0.006725	0.135398	0.000000	0.000000	0.000000	0.000000	2.500000e+01

Fig 1: Statistical description

The measure of the distance at which an object can be clearly discerned is the visibility that has been brought up from 0 to 140. Coming to the precipitation levels, though it has 0 readings at starting point and at different percent (25%, 50%, and 75%), the maximum it could reach was 25.

A heatmap uses a system where colors are used to represent different values in a graphical representation of data which can be shown in the figure 2 below. They are mostly used for showing user behavior on specific webpages or templates in various forms of analytics. From the heatmap considering the air pressure “Pressure(in)” and wind chill factor “Wind_Chill(F)”, they share a negative correlation considering the inverse nature of their relationship. Next, as the air temperature “Temperature(F)”, is increasing, its relative humidity “Humidity(%)”, is decreased as air can hold more water molecules making them negatively correlated to each other. The wind draws heat from the body as it increases and drives down the skin and body temperature making it much colder. This gives us a positive correlation between wind chill "Wind_Chill(F)”, and temperature “Temperature(F)”. For our next analysis, as the temperature increases, the relative humidity falls making the aerosols shrink, thus increasing the visibility that gives a positive relationship between “Temperature(F)” and “Visibility(mi)”. The vulnerability of vehicles is more if the distance affected by road accidents, traffic is more increasing the consequences in terms of injury and material damage. This

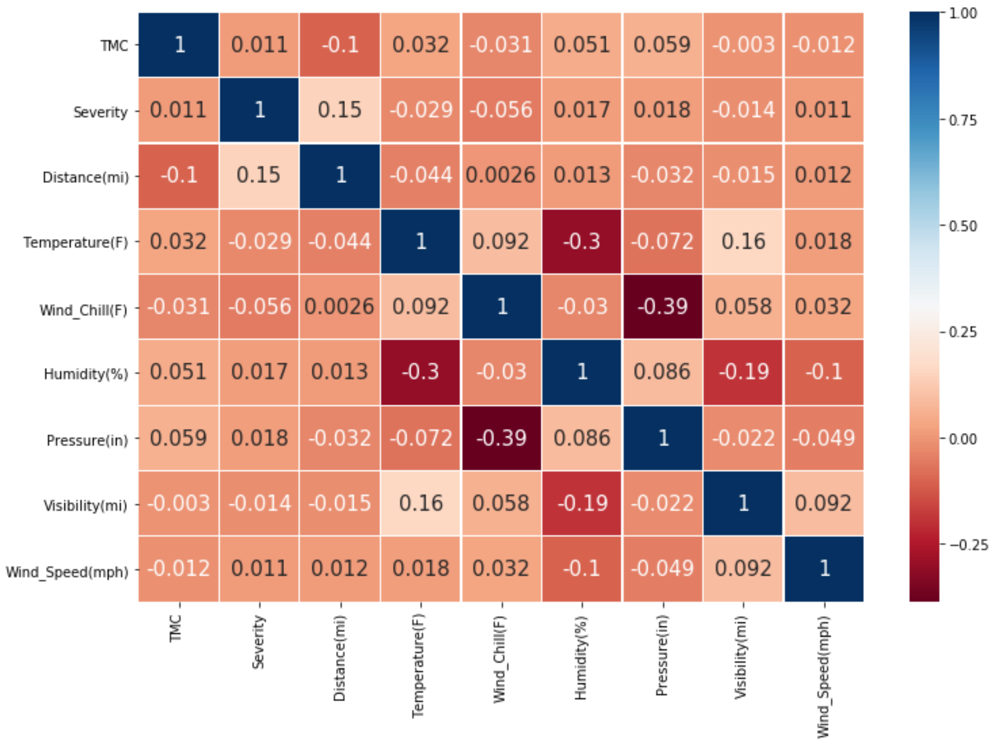


Fig 2: Correlation heat map

shows us that distance “Distance (mi) and severity “Severity” are positively correlated. The information on the severity in the dataset has been divided into four groups from 1 to 4. In the visualization below, information on the number of accidents that has severity impact 2 is highest, followed by 3, 4, and 1. From this, it can be understood that, since the data on severity 1 is low, the distance affected by road accidents is more because the number of accidents with the severity impact 2, 3, and 4 is high. The sources used to report road accidents globally are MapQuest, Bing, and MapQuest-Bing. Among these Mapquest reported the highest number of accidents and MapQuest the least. Coming to the visualization, (Number of Accidents by State), stated below, a Stacked Column chart has been used to plot the number of accidents occurred in each state. Through this it can be observed that more number of accidents occurred in the states, California (“CA”), followed by Texas (“TX”), and Florida (“FL”) and the least number of accidents in North Dakota (“ND”), followed by South Dakota (“SD”).

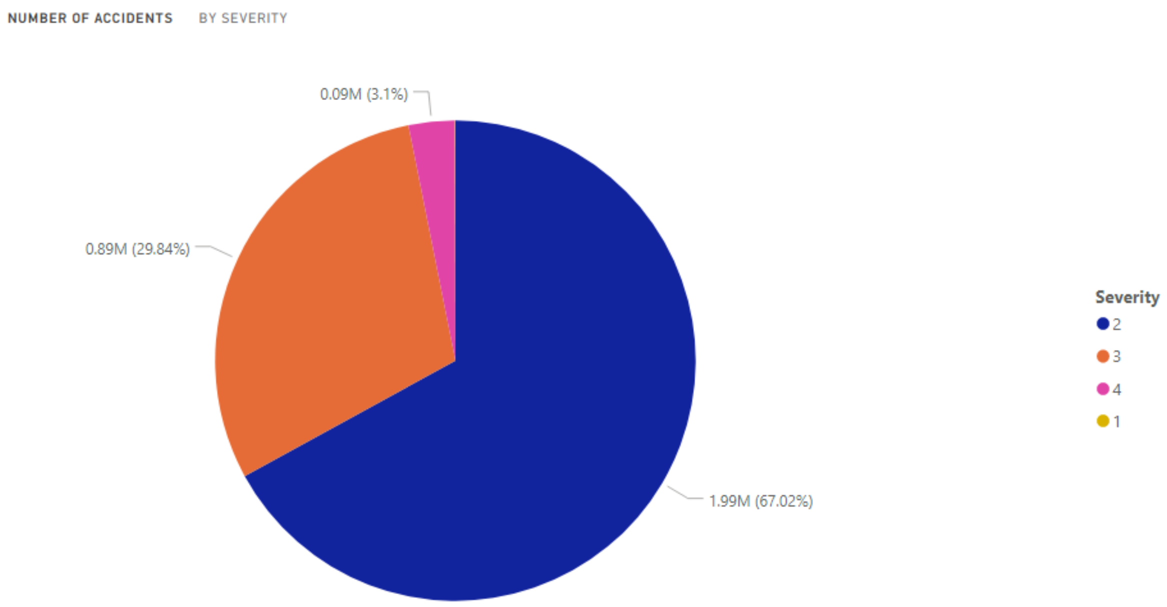


Fig 3: Pie chart showing the number of accidents by severity

When the records from the period 2015-2019 are considered, the year 2019 has the highest number of accidents and the year 2015 the least. Through this observation, it can be stated that the rate of accidents is gradually increasing from 2015-2019.

Hypothesis:

Now that we have done our analysis using EDA, we will get into the hypothesis for the dataset. For the first hypothesis, if we have bad weather conditions, then the rate of road accidents is high and if we have good weather conditions, the rate of road accidents is low. The second hypothesis is that the total number of accidents is high, if there are any interruptions in the traffic flow, such as amenity, bump, crossing, and junction. The last hypothesis states that the time period plays a major role in affecting the rate of accidents.

Results:

Power BI is a data visualization toolset that is customizable, giving a complete view of your business by transforming the data into rich visuals. The reason behind choosing this visualization tool is that it helps in analyzing data and sharing insights that allow the users to create personalized dashboards combining onpremises and cloud-born data in a single view. This allows users to monitor important data globally.

For the first and foremost hypotheses, we have information on climatic conditions such as wind direction, wind speed, wind chill, humidity, temperature, pressure, precipitation, visibility, and weather conditions, in the dataset. Among these, the factors, wind direction, wind chill, temperature, pressure, precipitation, visibility, have no relation to the rate of traffic accidents, hence they are not being included in our analysis. Therefore, the factors through which the analysis of the rate of these accidents is determined are wind speed, humidity, and weather conditions. To represent the first visualization, we have used Line and Stacked Colum Chart, plotted against Wind_Speed(mph) and the number of accidents as the ups and downs can be shown clearly in this chart. We are considering the total count of the column "ID" for the total number of accidents, as the ID for each accident is unique. From the visualization (Number of accidents by Wind_Speed(mph)), the maximum number of accidents occurred when the Wind_Speed(mph) is 0, and the maximum wind speed has been recorded at 30mph which has the least number accidents. Through the observation, it can be said that the number of accidents is gradually decreasing when the wind speed is increasing. A Stacked Colum Chart has been used to plot humidity against the number of accidents where severity has been plotted as the legend. As severity is plotted for the legend, we have used appropriate colors to differentiate the severity levels. The total number of accidents is high if the humidity is 0 and 100. From this, it can be inferred that the rate of road accidents is high for the extreme conditions of humidity. Another visualization has been shown for weather conditions against the number of accidents, using Treemap. The maximum number of accidents occurred when the weather condition is clear, followed by mostly cloudy, and overcast. For these weather conditions, most of them have severity levels 2 and 3.

I have represented the visualization using on-click interactivity which allows us to filter the information about a specific factor through just a click. The interaction can be performed when you click on any field of the Treemap. A Slicer has been used to select the various levels of severity.

Contrary, to the first assumption: if we have bad weather conditions, the rate of road accidents is high, and good weather conditions, the rate of road accidents is

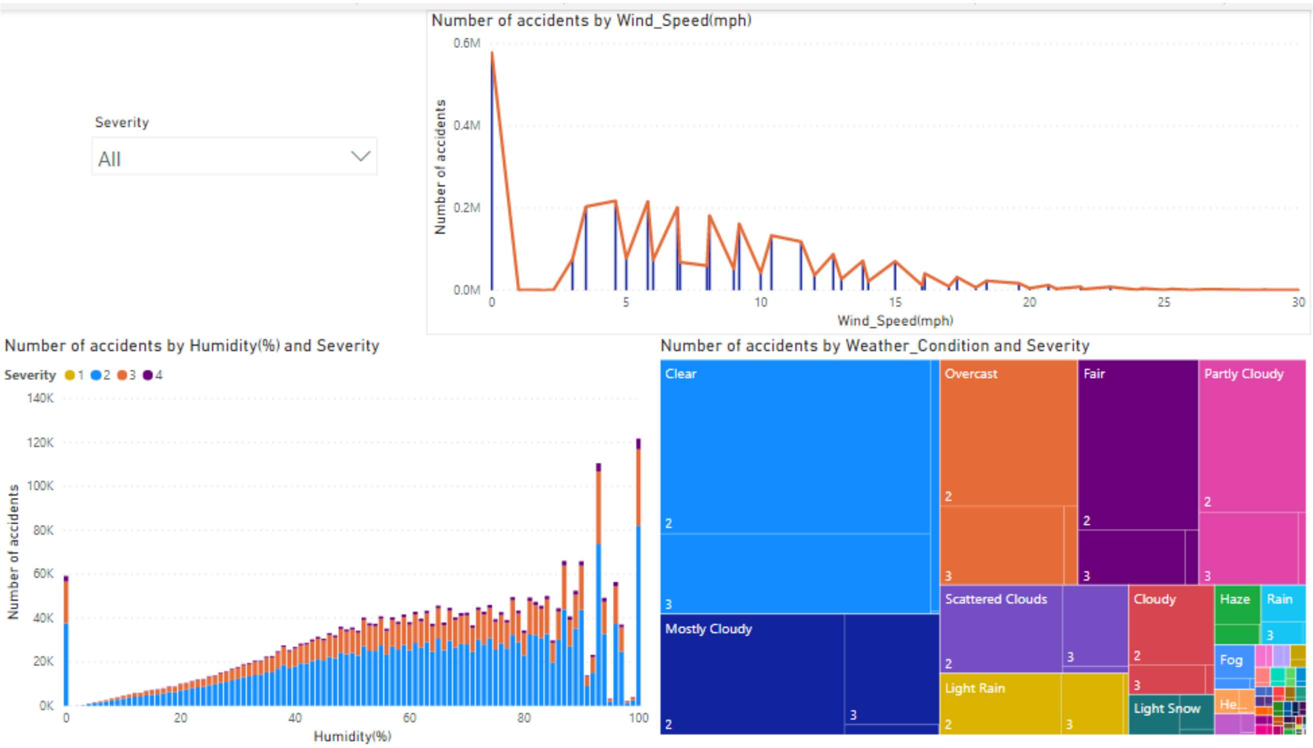


Fig 4: Visualization report for first hypothesis

low, the total number of accidents is high when wind speed is low which comes under good climatic conditions. Also, when the weather conditions are clear and fair (good weather conditions), the rate of traffic accidents is high. Supporting this assumption, the total number of accidents is high for extreme conditions of humidity. Overall, when all the climatic conditions are considered, good weather conditions have the highest number of accidents contradicting our first hypothesis.

For the second hypothesis we have data on traffic interruptions such as amenity, speed bump or hump, crossing, give away, junction, no exit, railway, roundabout, station, stops, traffic signals, and turning loops, in the dataset. Therefore, the factors through which the analysis of the rate of these accidents is determined are wind stop, crossing, junction, and traffic signals as the rest of them are not useful for further analysis. A Stacked Colum Chart has been used to plot the visualization against each traffic interruption and the total number of road accidents where legend has been used to differentiate the levels of severity. And each traffic interruption has been represented using a Pie Chart. From this report, it can be observed that the rate of accidents is high when there are no traffic interruptions.

This analysis is a contradiction to our second hypothesis that the total number of accidents is high if there are any interruptions in the traffic flow.

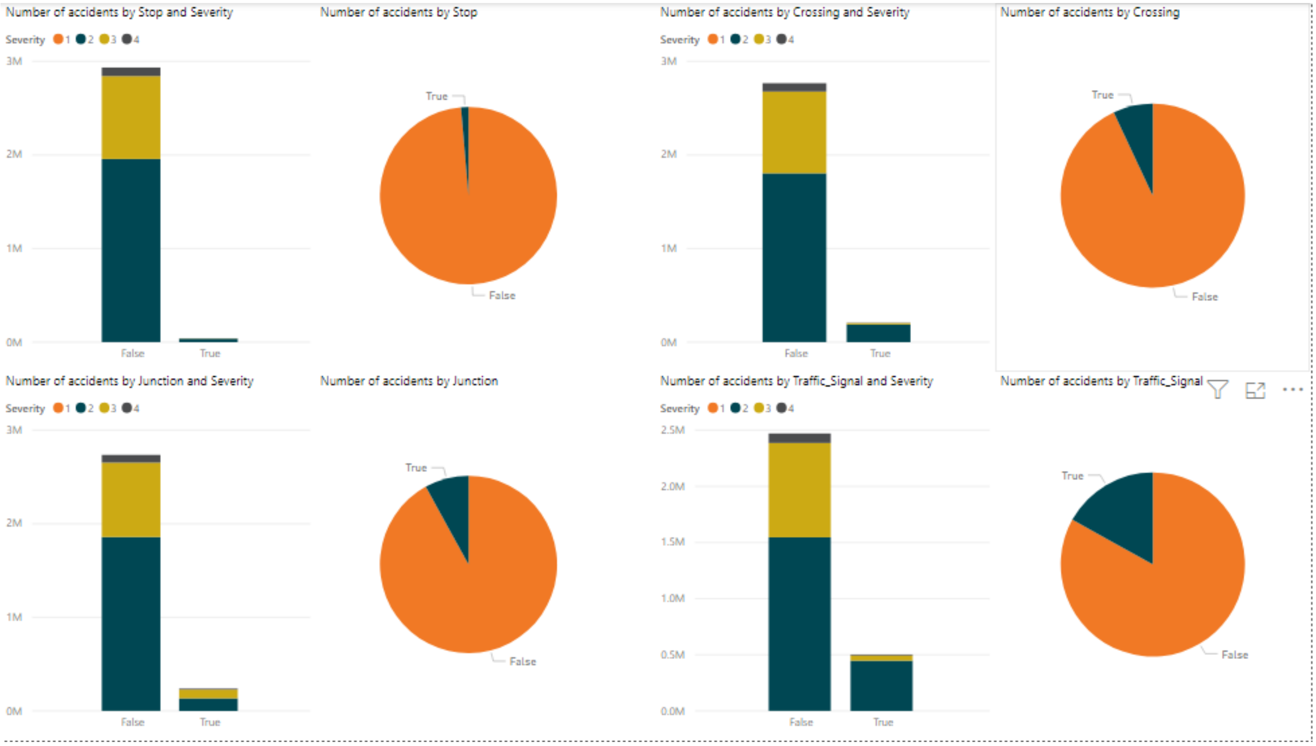
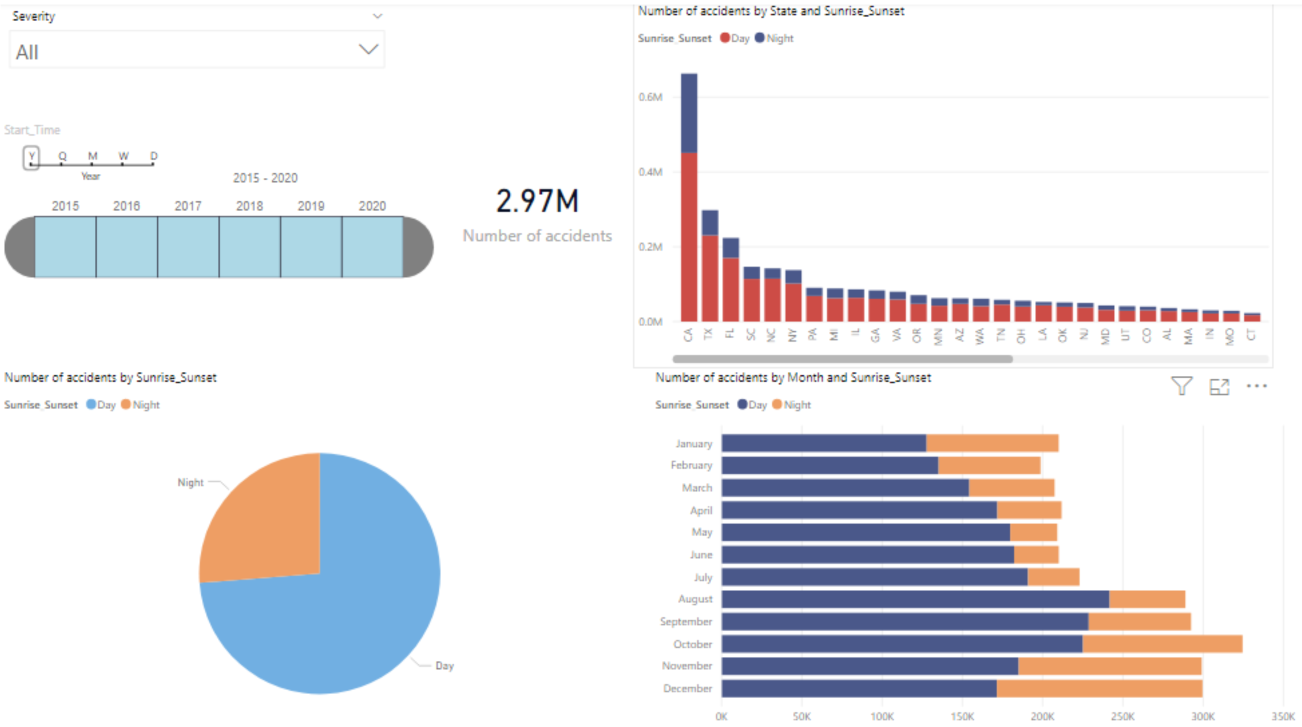


Fig 5: Visualization report for second hypothesis

For the final hypothesis, the first visualization (Number of accidents by State and Sunrise_Sunset) for the total number of accidents in the State, using a Stacked Column Chart where the field Sunrise_Sunset gives us the period (day or night) when the accident had occurred. This field (Sunrise_Sunset) has been plotted in a legend differentiating day from night using colors. According to the observation, the total number of accidents is high in California State, followed by Texas and Florida and low in North Dakota. When we notice the visualization, for all the states the total number of accidents, is high during the day. A Pie Chart has been plotted against the number of accidents by Sunrise_Sunset which is plotted in the legend to differentiate day from night. From this, it can be observed that around 2.2 million accidents occurred during the day and around 0.7 million accidents occurred during the night, out of 2.97million approximately. In the years from 2015 – 2019, 75% of the road accidents occurred during the daytime, approximately. A Stacked Bar Chart has been used to plot the number of accidents against each month and Sunrise_Sunset has been plotted in the legend to differentiate day from night. This gives that, for each year considered, the rate of road accidents is high in the month of October.

The visualization has been represented using on-click interaction. This feature allows us to filter the information about a specific factor through just a click. A custom visual, Timeline 2.1.1, has been used to filter the whole report using years, quarters, months, weeks, and days. A Slicer has been used to select the levels of severity. To display the number of accidents a Card has been used.



Severity

All

Fig 6: Visualization report for third hypothesis

From this report, it can be observed that the rate of accidents is high during the day and specific months of a year, which supports our third hypothesis: the time period plays a major role in affecting the rate of accidents.

Discussion:

An overview of the accidents such as their occurrence over years, which includes the number of accidents by each state, the best month, week, and day to travel, has been given by this project. It also focuses on the factors for the phenomenon of accidents like severity, time period, correlation, and traffic interruptions.

The first visualization has been built to show that the maximum number of accidents occurred when the wind speed is 0, with the maximum wind speed 30mph having the least number of accidents. Therefore, it has been said that the rate of accidents is gradually decreasing with the increase of the wind speed. The rate of road accidents is high in humidity in extreme conditions. When the weather condition is a clear maximum number of accidents occurred. From the overall results, the first hypothesis: the rate of road accidents is high for bad weather conditions and low for good weather conditions, has been contradicted as a good climatic condition, the wind speed was low, and good weather conditions are clear and fair, the accident rate was high. This assumption has been supported as the total number of accidents is high in humidity in extreme conditions. In contradiction to our first hypothesis, when all the climatic conditions are considered, good weather conditions have the highest number of accidents.

The second visualization focuses on the factors through which the analysis of the rate of the accidents is determined such as wind stop, crossing, junction, and traffic signals as the rest of them are not useful for further analysis. It can be observed from the report that when there are no traffic interruptions the rate of accidents is high. This analysis contradicted our second hypothesis where the total number of accidents was high if there are any interruptions in the traffic flow.

The final visualization observes that the total number of accidents is high in California State and low in North Dakota. From the visualization it can be noticed that all the states have a high number of accidents during the day, i.e., 2.2 million accidents occurred during the day and 0.7 million accidents occurred during the night, out of 2.97 million, approximately. Seventy-five percent of the road accidents occurred during the daytime for the years, 2015 to 2019, approximately. This gives that, the rate of road accidents is high in the month of October of each year. This gives us that the rate of accidents is high during the day and specific months of a year supporting our third hypothesis: the time period plays a major role in affecting the rate of accidents.

Future Work:

Results were not available for further research despite all these studies. Therefore, studies on road accident detection came into play. The total number of commercial and non-commercial vehicles on the road has meteorically increased by the improvement in transportation infrastructure and in-vehicle technology. The objective of the work is to create a system that automatically detects and notifies about traffic congestion in a timely manner in order to surpass the increasing number of casualties. This specific system can be achieved through the Internet of Things along with Vehicular Ad Hoc Networks.

We focused our discussion on proposing a platform that can showcase all the findings by each state and what period of time (day or night) is safe to travel, weather conditions, humidity, temperature, visibility, also if someone wants to go from one place to another, what are the difficulties one would face while traveling. This platform can help decide and provide solutions based on the issues of accidents that are faced by each state in the U.S.

References:

- National Highway Traffic Safety Administration. Traffic Safety Facts, 2000: Overview. DOT HS 809 329. Washington, DC: Department of Transportation, National Highway Traffic Safety Administration. https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100012882.pdf
- Najjar, A., Kaneko, S., & Miyanaga, Y. (2017). Combining Satellite Imagery and Open Data to Map Road Safety. Thirty-First AAAI Conference on Artificial Intelligence. https://scholarworks.lib.csusb.edu/cgi/viewcontent.cgi?article=2085&context=etd
- https://smoosavi.org/datasets/us_accidents