Car accident hotspots in the US and its causes

Abstract

Every year millions of accidents are reported all around the world out of which the US has plenty of cases noted. Many circumstances lead to accidents which include natural calamities or lack of safety measures taken by an individual. It's necessary to know the outcomes and analyse the causes of accidents to take possible actions to prevent such accidents from happening in future. The visualisation presents accidents occurring in 49 states of the United States with CA as the leading country with the largest number of accidents. Further, it is observed that most of the accidents appeared when the weather was clear and fair. Road features are observed to study the causes in deeper.

Data Collection

The data I used is publicly available on Kaggle as US Accidents: A countrywide traffic accident dataset (2016-2020) provided in CSV format having approximate 3.5 million records. The dataset is of size 1.24 GB and was captured using multiple traffic APIs. The latest version of data, Version 3, was updated in June 2020 while the next version is expected to be updated by December 2020.

The data has 49 columns and 3513617 records containing information of 49 states of the country from February 2016 to June 2020. The attributes have information related to accident severity, weather conditions, source of Accidents, Accidents in day or night, etc.

According to me, all the three aspects of big data are part of my dataset. Firstly, it has volume since it contains large records with 1.24 GB size. Secondly, it also contains velocity as the data is found in version 3 and is updated after every 6 months. Finally, the variety of data is enough to give conclusions on Accident analysis with a lot of informative attributes.

Data Exploration, Processing, Cleaning and/or Integration

Data Exploration

Firstly, I explored data by reading all the descriptions of each attribute. This helped me in creating imaginary visuals in order to make a note of necessary columns that I will be using to design graphs in the next step. The three visualization questions that I shortlisted after exploring data were

1. What is the state-wise accidents across the country and what time of the day did the accidents occur?
2. Does the weather affect the accident numbers?
3. What is the severity of accidents in the state with the highest number of accidents?

Data Cleaning and Processing

The raw dataset had many missing values and null rows present in it. Before processing missing values and eliminating null rows I discarded those columns which were not required in my visualization process.

Attributes like Temperature, Humidity, precipitation and visibility with float data types had many null values so I replaced them with their median values respectively. After filling missing values, there were very less null value rows left hence I eliminated those rows. The final filtered data had approximated 3.4 million records after the filtration of data.

Visualisation

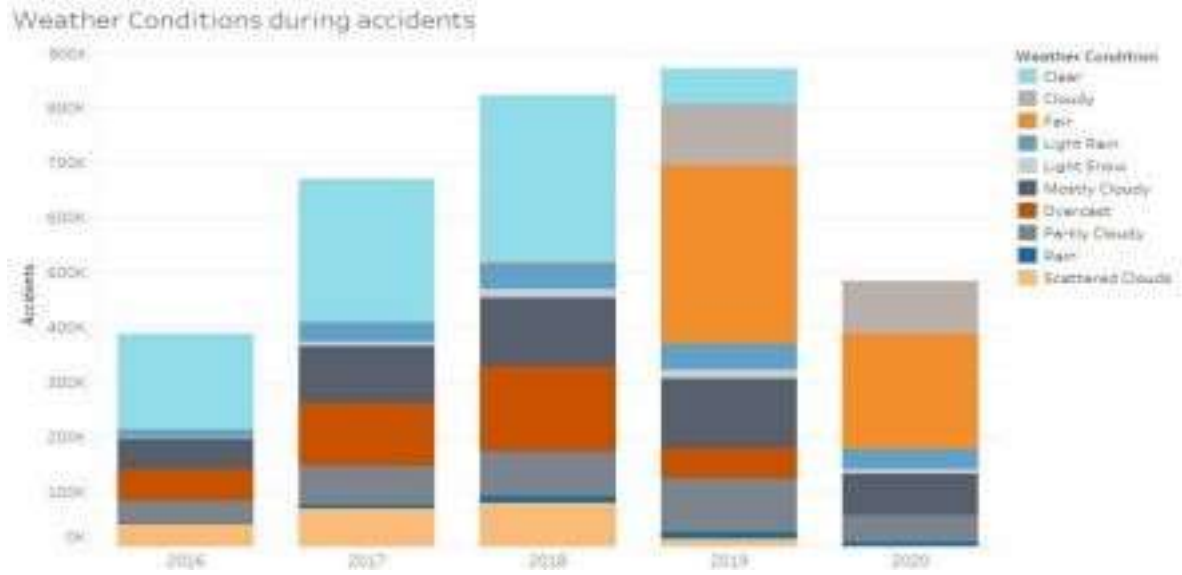
The first chart is an animated map describing the accidents of each 49 states. The Pie charts over every state indicates the percent of total accidents appearing in Day and Night. The animation shows that from 2016 till 2020 CA, TX, FL are the leading states in US with highest accidents reported every year specially CA alone reporting 1 million plus accidents each year. It is clear that most of the accidents occurred in Daylight.

Choice of chart: I wanted to compare states on the count of accidents from lowest to highest accidents from 2016 till 2020 using an animation. Here Bar charts could have been an easy option but I chose a map chart because showing 49 states in one bar chart would have made it difficult to view. Also, Bar chart

would require a lot of scrolling if new data according to future years is updated. Hence Map is the best visualization to present State Wise number of accidents. Pie charts over each state made it easy to give a clear understanding of accidents happening in day or night since Pie charts are easy to visualize and derive conclusions when there are only two measures to analyse.

Design choices: I choose red to indicate the count of accidents based on states since red defines danger. So the darker the red, the worse the hotspot area it is for car accidents. I choose cool colours like blue and grey to indicate day and night respectively as these colours do not strain eyes when used over warm colours such as red.

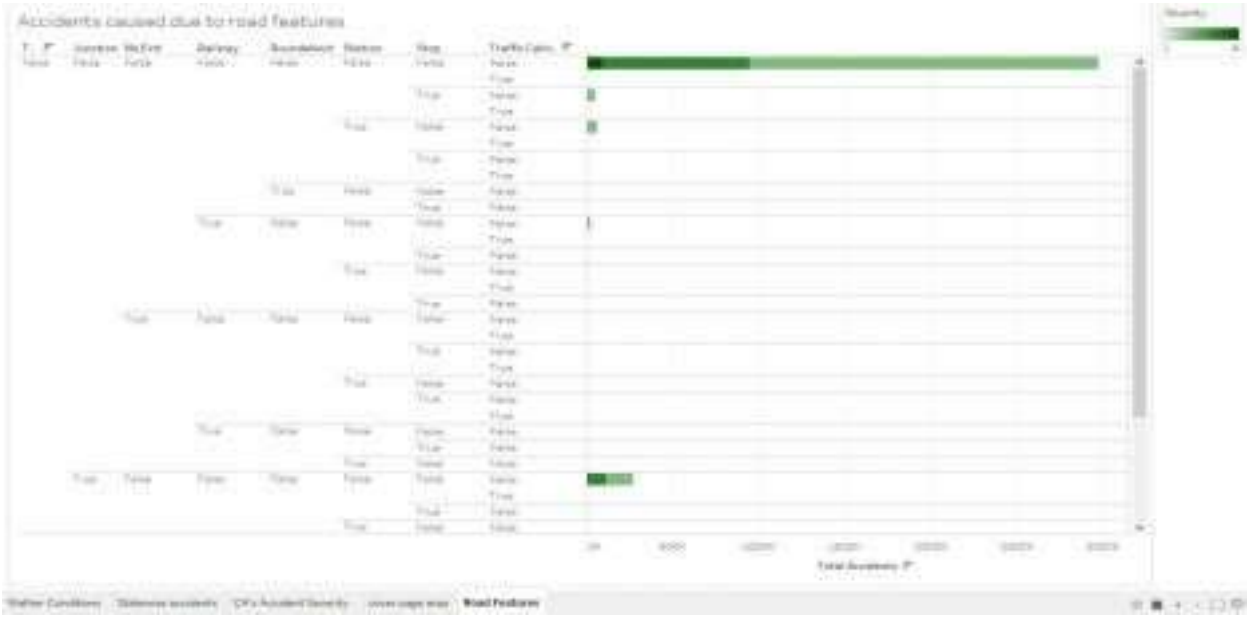
The Next visualizations try to give an overlook of the causes such as weather conditions or Road features that might lead to accidents in above states.



The stacked bar chart given above demonstrates the weather conditions when the accidents occured. It points out that most of the accidents happened in cloudy and clear weather from 2016 till 2018 while 2019 and 2020 has most accidents in fair weather. This conveys that bad weather did not affect the high numbers of accidents. This might be because drivers are more likely to drive slower in rains and snow.

Choice of chart: I used a stacked bar chart as it helped in visualizing all the weather categories at the same time making it easier to blame which weather caused more accidents in given years.

Design choices: I preferred to use a colour-blind palette which is inbuilt in tableau to show stacked data so that nobody finds difficulty in comparing weathers with multiple colour legends.



Now let's look at the road features along with the severity of accidents. The above Gantt chart shows that most of the car crashes were due to junctions, stop, railways and station. Junctions and railways caused large number of level 4 severe accidents.

Choice of chart: I chose Gantt chart since it was less complex to read Boolean values of every road features and check the severity levels.

Design choices: I kept the chart as simple as possible to read since it has lot of data to read and understand. Also using one shade to visualize the range of high and less severity of accidents is more preferable than using different colours,

Tools and libraries used

Jupyter Notebook was to write Python codes. Python was used to do all the cleaning and processing.

Tableau was used to make all the visualization graphs and maps.

Conclusion

The data was quite informative to think over various interesting visualizations. I spent a lot of time exploring a cleaning tool for processing the raw dataset. I explored spreadsheets and python using jupyter notebook. There was a little difficulty in data loading since my data file was very huge to get loaded at once in spreadsheet and hence required breaking of file in parts. Since there were large number of missing value fields in the dataset, cleaning in spreadsheet was a bit of a long process for me hence I finalize python for cleaning and processing data in one go.

For the map chart, I wanted to distinguish states with more colour options in red shades since the states below 50k accident records appear to be in same colour. This could have been improved by studying more colour patterns and effects in Tableau. I also used dual axis to make the chart look interactive by representing the state wise accidents along with the day and night percentage of accidents.

To study the weather conditions, stacked bar charts were the best option that helped me in giving clarity of categorical data. It helped me giving clear visuals of which weather caused more accidents. Finally, to find more causes of car crashes I used Gantt chart to give which gave me a structured representation of road features that lead to more severe accidents. All the visualizations helped me in concluding that CA has highest number of car accidents while all over the country most of the Accidents took place in Daylight and clear weather. Also, the road junctions, stations and railways are the main causes of severe accidents.

References

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2019)

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International
Conference on Advances in Geographic Information Systems, ACM, 2019.