Traffic collisions in California over the years

Abstract

A traffic collision also referred to as a motor vehicle collision, or car crash occurs when a vehicle collides with another object that can be a vehicle, pedestrian, animal, or other stationary obstruction, such as a tree. Traffic collisions [2] often result in injury, disability, death, and property damage as well as financial costs to both society and the individuals involved. Road transport is the most dangerous situation people deal with daily, but casualty figures from such incidents attract less media attention than other, less frequent types of tragedy.
The story conveyed about the traffic collisions in California over the recent years giving a time series aspect through visuals. Firstly, the visualization highlighted the number of collision cases that occurred over the different segments of time i.e., year, quarter, month, and day. Secondly, vehicle code violation turned to be a primary reason for the collision, and the majority of cases involve rear-end collisions.
Furthermore, the number of different casualties i.e., pedestrian killed, a bicyclist killed, and motor-cyclist killed occurred in collisions over time. Moreover, in no surprise, due to the COVID-19 pandemic outbreak, there is a steep decline in the trend of traffic collisions this year.

Dataset

My data is provided in SQLite database object instance by the Statewide Integrated Traffic Records System (SWITRS). The dataset is available on Kaggle developing platform [1]. In my opinion, the three aspects of big data are present in my data in the following ways.
Firstly, it has velocity: the traffic collisions dataset is updated often by adding the new cases reported. The dataset is serving as a data lake having information from 2001 till current date collision records. Along with the fact that the repository is maintained for information regarding the collision to the current date, this fulfills the volume aspect.
Also, the dataset is from SWITRS itself i.e., is an accurate and trustworthy site to port the data. This covers the veracity aspect by having data available from reliable sources. From the data I collected, I have the information from three tables: collisions, parties, and victims. There are a total of 2178113 rows, 4345454 rows, and 1775443 rows respectively in these three.
Moreover, the repository size 5.78 GB is maintained in SQLite and contains 3 tables, presenting various information of collision cases.

- collisions: Contains information about the collision, where it happened, what vehicles were involved.
- parties: Contains information about the group's people involved in the collision including age, sex, and sobriety.
- victims: Contains information about the injuries of specific people involved in the collision.

Each of the above tables is at the granularity of case id i.e., unique identification for the collisions reported to date.

Data Exploration, Processing, Cleaning and/or Integration

Data Processing

The data was collected from Kaggle and exported to CSV files with the help of python in Spyder IDE. I have loaded the SQLite3 package to set up a connection with the SQLite database. SQL queries were written using the connection object to pull the information for collision and integrating with other tables to reduce the data for limited case ids. The source data is available from 2001 and I used a subset of recent 5 years of data. Now, each SQLite table is converted to a python data frame object.

Data Exploration, Cleaning, and Integration

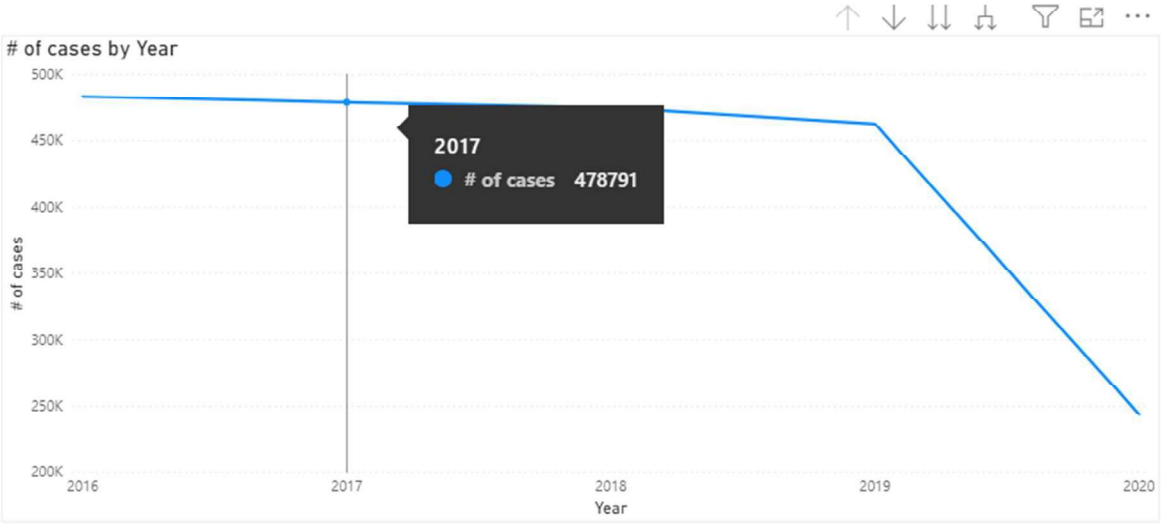
I have exported the data frames to comma-separated (CSV) files which are further used for graphs/charts. Power BI tool was used for data cleaning, reductions & transformations, and data visualization. Post loading the CSV in PowerBI, I have dropped the unwanted set of columns from all the tables and remove the rows having nulls/blanks from the required set of fields.

In PowerBI we have an ER model representation of the tables used along with joining keys and cardinality; made the changes and updated the cardinality between the tables. Also, the incident dates are transformed to a hierarchal structure to provide the information at a year, quarter, month, and day level.

The attributes chosen are the key attributes to understand the data and to answer the question like how many cases happened over time? What are the type and reasons for those collisions? What does the headcount of casualties happen with pedestrians, cyclists & motor-cyclist?

Visualisation

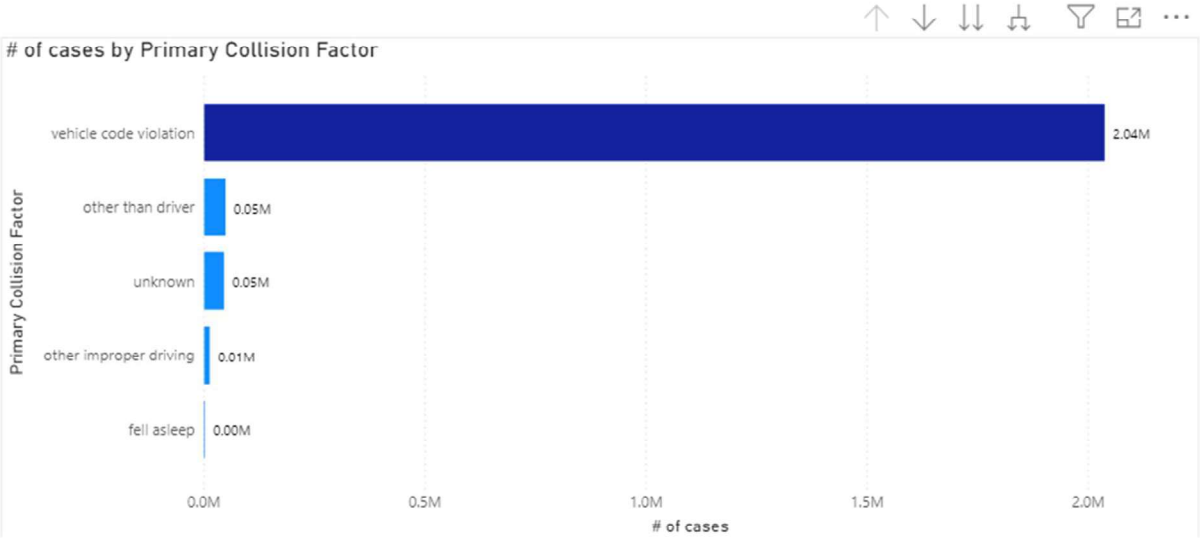
The first one is an interactive drill-down histogram line chart to show the distribution of the number of collision cases over time.



The histogram line chart is used here because I need to represent the distribution of an attribute over time. The periodical data along with hierarchal collision date help in seeing the data at a different granular level. Also, the chart can specifically be drill-down to have information filtered

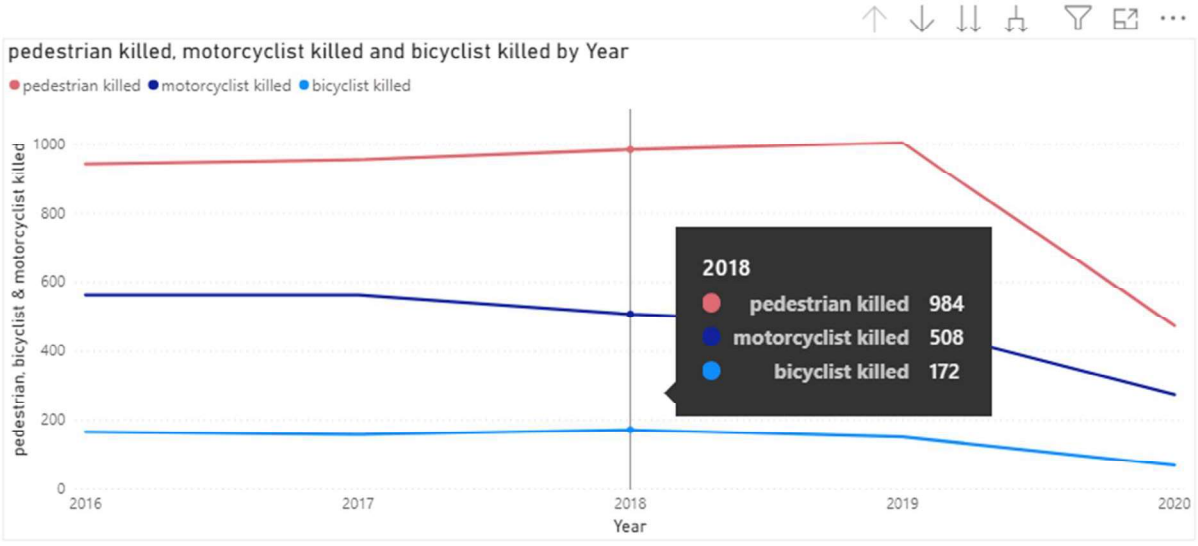
from the above level example, we can refer to see quarter level distribution data for a certain year. I chose blue from dark-color palette as having single data value.
I made the changes to graph format by removing the gridlines, renaming the titles and both axis titles along with the size of both axis data points to have clear readable fonts.

The second one is an interactive drill-down clustered bar chart to show the different reasons for the collisions and can be drilled to check on the type of collisions for the same.



The clustered bar chart is used here because I need to compare the number of cases in different collision factors. I made the changes to graph format by removing the gridlines, renaming the titles and both axis titles along with the size of both axis data points to have clear readable fonts. The data labels are also to check clearly the factors with small values. A vehicle code violation is a prime factor for collision, so I have highlighted that bar with a darker shade of blue.
Also, the chart can specifically be drill-down to have information filtered from the above level example, we can refer to see the type of collisions for a single collision factor.

The third one is an interactive drill-down line chart to compare the number of pedestrians, bicyclists and motor-cyclist killed over time.



The line chart is used here because I need to compare killed counts in different categories over the time period. The hierarchal collision date help in seeing the data at different a granular level. Also, the chart can specifically be drill-down to have information filtered from the above level example, we can refer to see month level distribution data for a certain year and a certain quarter.
I made the changes to graph format by removing the gridlines, renaming the titles and both axis titles along with the size of both axis data points to have clear readable fonts. The pedestrian killed counts are highest amongst the three, so I highlighted it with a red color and giving another color by reducing the tone of color for other categories. These categories are clearly available in the legends for reference and clear distinction between each other.

Moreover, all the charts are interactive with each other, and the selection of one data point leads to filter in other charts. The charts are referenced and created considering the best type of chart as per data [3]. The tool used for data processing is Spyder IDE (Version 3.3.3) and the cleaning, integration, and transformation along with visualization are carried out in Power BI desktop (Version: 2.83 64-bit (July 2020)).

Conclusion

The biggest challenge that I face is with the data pre-processing phase. There were a lot of unidentified and missing values. Post cleaning the charts turns easier to plot. I spend time implementing different interactive aspects of the graphs.
Considering the first chart, I tried with other graphs like the heatmap and geospatial map with the latitude and longitude position to identify the targeted spots of collisions but the chart was getting difficult to analyze and interpret from the end-user perspective. The periodical data is better represented using the above chart and the major outcome that came from the graph is a steep decline in the number of case records in the current year due to the COVID-19 outbreak. Even for the major holiday time, you will see a reduction in the number of cases.
In the second chart, the clustered bar chart seems to be the best possible option considering different categories compared on a value. Regarding the chart format having the color saturation over the bar based on values can also be a good option to visualize the chart. The outcome came out to be as expected, the most common factor of collision is violating the traffic rules. And, the rear-end collision came out to be with the highest number of collisions happened.
For the third chart, I wanted to have six attributes as data values instead of three to include the injured counts. However, although obtained and can represent six values the graph would turn complicated and hard to read. Moreover, I also tried to have some dynamic filter implementation to switch between the killed and injured data categories, not worked at last. The outcome came from the graph is mostly victims that got killed are pedestrians.

References

1. Gude, A., 2020. *California Traffic Collision Data From SWITRS*. [online] Kaggle.com. Available at: https://www.kaggle.com/alexgude/california-traffic-collision-data-from-switrs [Accessed 16 December 2020].

2. En.wikipedia.org. 2020. *Traffic Collision*. [online] Available at: https://en.wikipedia.org/wiki/Traffic_collision [Accessed 16 December 2020].
3. Guy, M., 2020. *Types Of Charts And Graphs: Choosing The Best Chart*. [online] My Market Research Methods. Available at: https://www.mymarketresearchmethods.com/types-of-charts-choose [Accessed 18 December 2020].