California Traffic Collision Analysis

Manisha Paliwal
Department of Applied Data Science
San Jose State University
San Jose, United States
manisha.paliwal@sjsu.edu

Sonal Sharma
Department of Applied Data Science
San Jose State University
San Jose, United States
sonal.sharma01@sjsu.edu

Praphul Kenkere Omkarmurthy
Department of Applied Data Science
San Jose State University
San Jose, United States
praphul.kenkereomkarmurthy@sjsu.edu

Chidroop Sagar
Department of Applied Data Science
San Jose State University
San Jose, United States
chidroop.sagar@sjsu.edu

Harshitha Ramesh
Department of Applied Data Science
San Jose State University
San Jose, United States
Harshitha.ramesh@sjsu.edu

***Abstract*— Road accidents are one of the leading causes of deaths around the world. National Highway Traffic Safety Administration reports state that the automobile traffic crashes in 2020 were increased by 7.2% when compared to the year prior. The increase in use of automobiles has also increased the rate of traffic on the roads of California. This project focuses on studying the road traffic accidents specifically in the state of California from the year of 2019 - 2020. This project will scrutinize the dataset from Kaggle repository as recorded by California Highway Patrol and perform data analysis using Jupyter notebook and Tableau to draw inference on the changes on rate of accidents during the 2019 and the Covid phase in 2020. The analysis focuses on comparison of traffic collisions between years 2019 and 2020. Results show that fatal and non-fatal collisions were high during January 2019 and reduced during the initial months of the year 2020. Most collisions were caused by passenger cars for 2019 and it significantly reduced by 37% in 2020.**

Keywords*— *California Highway Patrol, Traffic collisions, Accident risk

I. INTRODUCTION

The increase of automobile demand has also increased the traffic rate across many countries. The increase in automobile demand has also increased the traffic and the traffic collision rate compared to the last decade. Over 1.2 million individuals die every year on the world's streets, and somewhere in the range of 20 and 50 million endure non-fatal injuries. In 2019, California's mileage death rate was 1.06 fatalities per 100 million miles traveled. That's according to the Global Traffic Scorecard released in March 2020 by INRIX, a data analytics company that studies how people move around the world – San Francisco, California was rated 7th among the most congested cities in the U.S. in 2019. The collisions are caused due to various reasons such as alcohol or drug consumption, cellphone in use, motorcycle, bicycle, pedestrian etc. During the pandemic in 2020 nationwide lockdown had restricted the movement of traffic. In this study, we are analyzing the dataset of Traffic collision rate of California from the Kaggle which was provided by California Highway portal records to the author. The data was collected, clean, manipulated, tabulated, and then analyzed. The analysis shows the accident prone regions of the country and the comparison of the fatal and non-fatal collisions in year 2019 and 2020.The significant findings from the analysis are: (a) most of the fatal and non-fatal accidents occurred in January 2019 (b) a more substantial number of deaths are from drivers in the 20-50 age group; (c) Most of the fatal collisions have occurred during the cloudy weather followed by rainy weather (d) about 13% accidents are attributed to drunk driving. (e) Passenger cars contribute to the highest number of collisions in 2019 and 2020. 22,464 collisions occurred in 2019 which dropped to 8320 in 2020 during COVID (f) Top 3 traffic violations leading to collisions were : Not following Traffic guidelines(11%), Unsafe speed (22% approx.) and Improper turns(17%).

II. BACKGROUND

In this section, we first introduce what a database is and what our database consists of.

A. A database is an organized collection structured information that is stored electronically in a computer system. Connolly and Begg define database management system (DBMS) as a "software system that enables users to define, create, maintain and control access to the database". The DBMS manages incoming data, organizes it, and provides ways for the data to be modified or extracted by users or other programs.

B. Data Analytics refers to the process of analyzing the raw data and finding out conclusions about that information. Data Analysis in studies and research helps to reduce large dataset into a story and interpreting it in forms of graphs to derive insights. Descriptive, Diagnostic, Predictive, Prescriptive are the four basic types of data analytics.

Based on our dataset we are using are the Products, Descriptive, Diagnostic, Predictive, Prescriptive are the four basic types of data analytics. The California Traffic collisions

dataset used in our analysis is from Kaggle repository which is collected from the California Highway Patrol records by the author. Our dataset is refined for the records of traffic collisions during the year 2019 and 2020.

III. RELATED WORKS

A lot of studies has been in the field of Data Analytics. This study determines the risk values and detail this process using the NSW traffic accident database. They determine the effects on risk of particular vehicular behaviors such as speed and headway can be calculated and use these results to modify vehicle behavior in real-time to maintain a predefined risk limit. The results show that it is possible to reduce the accident rate among vehicles while at the same time increasing road network throughput by exploiting the variation in risk between vehicles [1]

Another study published in ICBIR presents the overview of factors influencing the road traffic accident severity and also reviews the techniques that frequently used in previous studies such as logistic regression, power model, and etc. From the review of literature, the mostly mentioned factors that are found to be significant to road traffic accident severity are speed of vehicle traveled, followed by human characteristics. Other factors that are found to be significant are vehicle types, weather, alcohol consumption, driver's fatigue, and etc. [2]

A study in *IATSS Research* presents a collision reporting system based on TCP/IP protocol. The system is composed of three parts, vehicle collision sensor terminal worked on car cigarette lighter, application used on smartphones and remote server based on Lab VIEW software. At the end of the paper, remote control and alarm tests are carried out. A detailed evaluation of the proposed system and collision reporting demonstrates the suitability as Vehicle Collision Reporting System for human drivers. [3]

IV. PROPOSED DATA VISUALIZATION

A. *Data Source*

The California Traffic collisions dataset that we have used is from Kaggle repository which is collected from the California Highway Patrol records by the author. Our dataset is refined for the records of traffic collisions during the year 2019 and 2020. This dataset consists of 3 tables: Collisions, victims and parties. The dataset size of original dataset is 10 GB. Below are the statistics of columns in each table: Collisions: 75 columns, 9424334 rows, Parties:32columns,18669166 rows, Victims: 12 columns, 9639334 rows.

B. *Data wrangling*

As part of the project scope, we selected traffic collisions that occurred in year 2019 and 2020. There are ~ 5k cases of traffic collisions. The data set has 49958 unique cases and their related collisions, parties and victim's data, which is not in the right format to start the analysis. After data wrangling, the dataset size is 46.6 MB. It contains 9 entities as below: Collisions: 35 columns, 49958 rows, Parties: 28 columns,100674 rows, Victims:12 columns, 45871 rows, County: 2 columns, 47 rows, Collisions location: 7 columns, 41996 rows, Road Condition: 3 columns, 46 rows, Vehicle Type: 2 columns, 16 rows, Violation: 3 columns, 255 rows, Weather_Effect:3 columns,30 rows.

C. *Normalization*

The initial California traffic collision dataset available contained three entities - Collisions, Parties and Victims in deformalized form. As part of this project, we have normalized the entities up to third Normal Form modelling it into a snowflake schema.

D. *Creation of procedures and triggers*

We have created 2 triggers and 1 procedure with this dataset. The first trigger gives the aggregated count of party involved in a particular collision briefly and gives the aggregated count of victims in a particular collision briefly. The stored procedure Compare_Collisions_2019_2020 will be further used for visualization.

E. *Queries to find the distinct counts*

F. *Connecting with AWS*

In this project we have uploaded MySQL Database in AWS-RDS (Amazon Web Services – Relational Database Services), further we were able to connect AWS RDS with python to access the AWS RDS database and perform SQL queries

G. *Visualization*

Comparison of California fatal collision for 2019 and 2020.

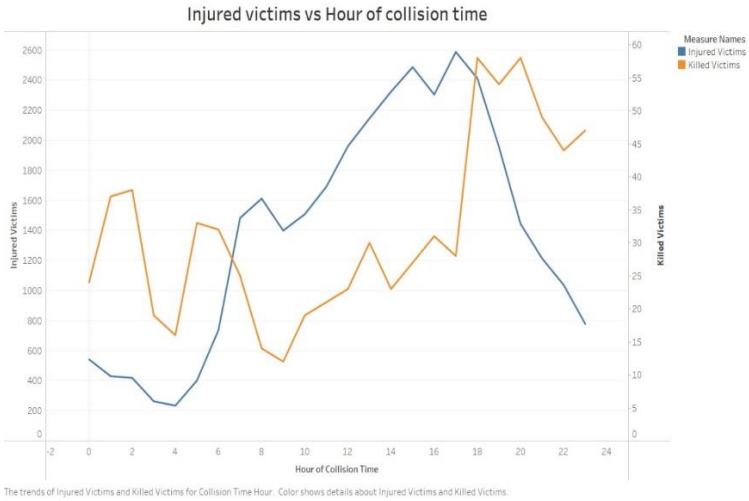

FIGURE 1: GRAPH REPRESENTING THE INJURED VICTIMS VS HOUR OF COLLISION TIME

Above line graph illustrates the correlation between number of injured and killed victims' and time of the day. Units measured are in hours (0-24-hour format).

It shows the dependency and correlation of injured and killed victims. Overall, the number of victims injured and killed is at the minimal till the peak afternoon. And, following that is a steep increase in the number of injured victims. On the other

hand, the number of killed victims is at the peak from evening 6pm till midnight.

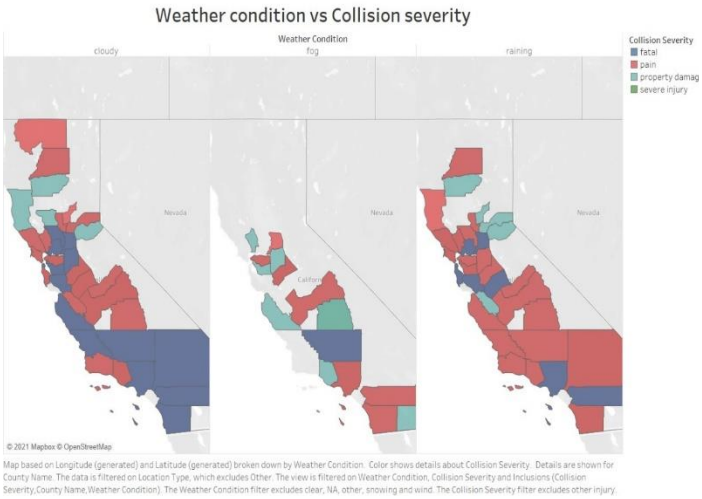

FIGURE 2: GRAPH SHOWING WEATHER CONDITION VS. COLLISION SEVERITY

This figure illustrates the collision severity with respect to changes in weather.

The major contributors for weather related accidents are cloudy, fog and rain. Highest fatality rate can be witnessed when it's cloudy. Whereas, it's the least when the weather is foggy. Highest pain is witnessed when it's raining and it's the least when the weather is foggy. This is a clear indication that foggy weather is comparatively better than cloudy and rainy weather condition.

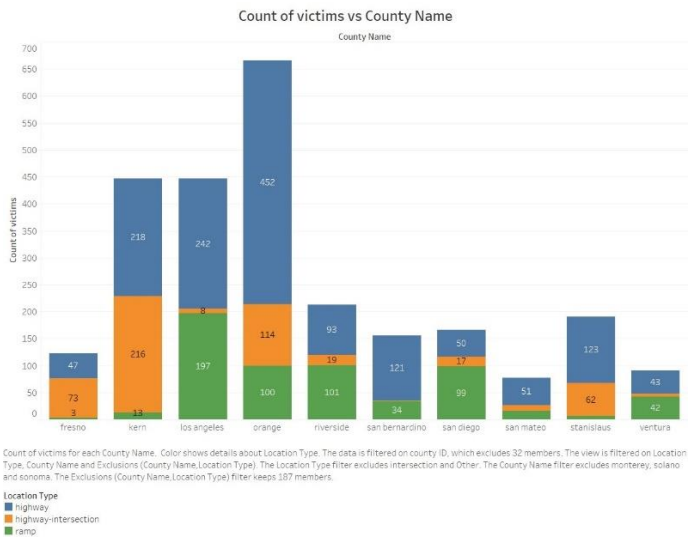

FIGURE 3: GRAPH REPRESENTING ALCOHOL INFLUENCE VS. TIME AND TYPE OF COLLISION

The above graph gives a pictorial description on the total number of victims who have been in a collision in different counties in the State of California for the years 2019-2020 in different locations. In the High-intersection roads Kern has the largest count (216). Kern has the most victims in both Highway and Highway intersection when compared to most of the other counties.

Roads with Ramps also have a good number of victims with Los Angeles having a count of (197), followed by Orange county (100). These two counties have a fairly large number of victims in both the Highway and Ramp roads. Fresco and Stanislaus have the least Ramp road victims count. We can infer from the graph that most of the accidents occur in large counties such as Los Angeles, Kern and Orange with Highway and Highway Road Intersections

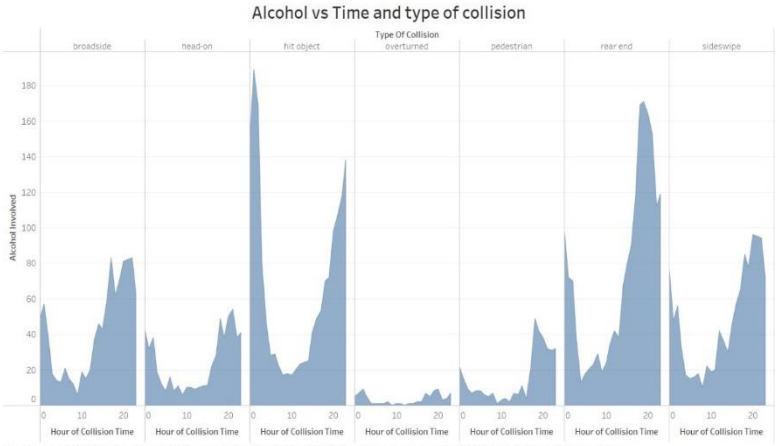

FIGURE 4: ALCOHOL INFLUENCE VS. TIME VS. TYPE OF COLLSIONS

The above graph gives a pictorial description on the type of collision and the time of occurrence when the person is under the influence of alcohol. We can infer from the graph that the most common type of collision is Hit the object collision and the most common time for an accident to occur while the person is under the influence is usually during the night time and during major holidays.

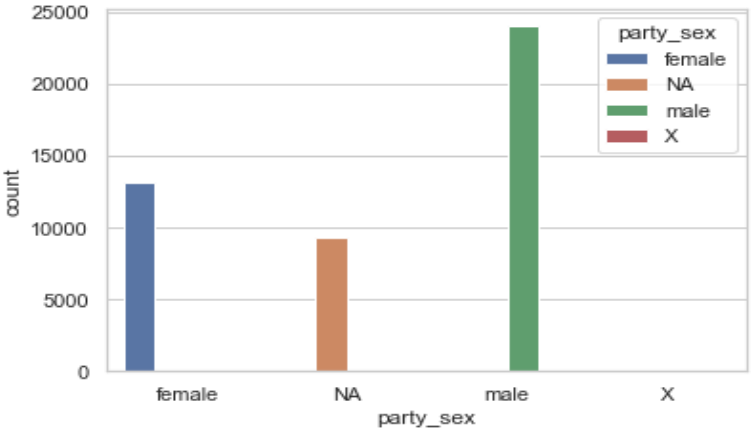

FIGURE 5: GRAPH SHOWING GENDER COUNT OF PARTIES

The above count plot done in Python visualization gives us insights on how many people were involved in the collision and their gender either Male or Female and NA is the unspecified data in the Dataset which means their gender was not specified. It is clear that Males have a larger count with a number of almost 25000 and females have a count of 12500.

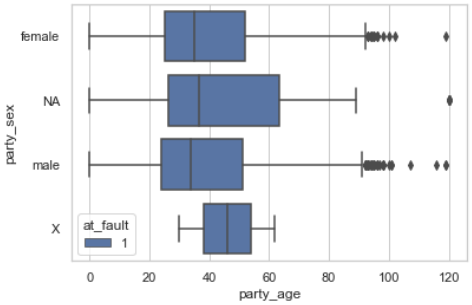

FIGURE 6: GRAPH SHOWING AGE GROUP VS. PARTIES GENDER

The above boxplot gives us insights on the party age, party sex based on Female or Male and Na. From the graph we can say that the mean age for male victim at fault for causing an accident is 35 and the mean age for the female victim causing an accident is 37.

From the graph we can say that the age group that are at fault is mainly between early 20's and 40's for both the genders. Na is the not specified gender which tells us that there is a large number of people between the age group of 30 and early 60's who are at fault for the accident and their gender has not been specified while collecting data.

We can infer from the above two graphs that the total number of men victims at fault are the highest when compared to the female victims at fault.

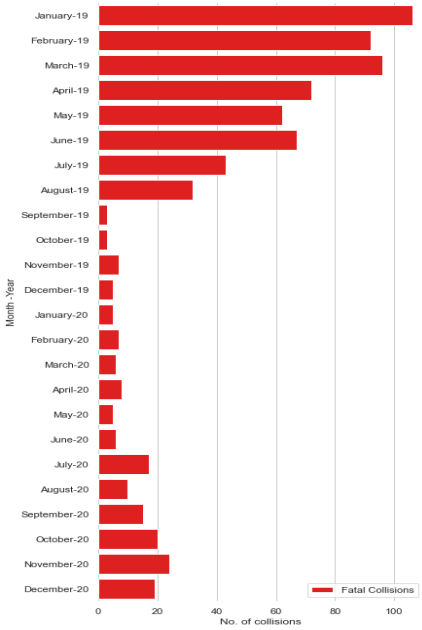

FIGURE 7: GRAPH REPRESENTS NUMBER OF FATAL COLLISIONS FROM JANUARY 2019 – DECEMBER 2020

The above bar plot provides insights on how many fatal collisions takes place from January 2019 to December 2020. It is clear that the number of fatal collisions pre-covid was more as there was more movement amongst the people of California. January-2019 has the most with a count of almost greater than 100 fatal collisions. But as the pandemic started to get worse the movement within the state reduced so did the total number of collisions in the State as well. The months from September 2019 till July 2020, from this we can infer that the movement of people reduced and gradually so did the collisions. But soon as they started to ease the lockdown rules, the rate of collisions increased as people started to travel. The fatal collision cases include sever injuries and sometimes even loss of life.

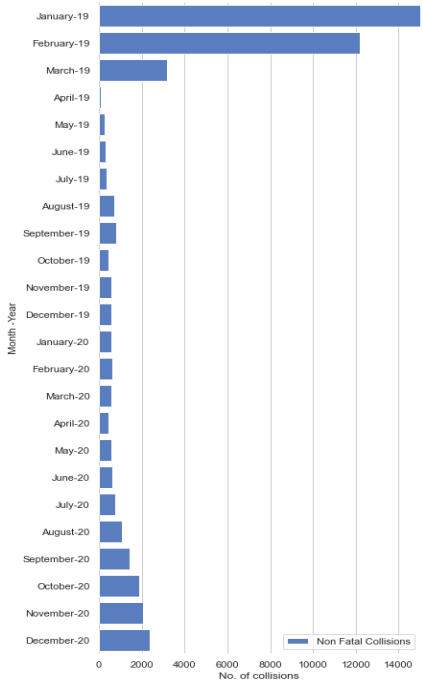

FIGURE 8: GRAPH SHOWING NUMBER OF FATAL AND NON FATAL COLLSIONS FROM 2019 JANUARY – DECEMBER 2020

The above bar plot provides insights on how many non-fatal collisions takes place from January 2019 to December 2020. It is clear that the number of non-fatal collisions pre-covid was more as there was more movement amongst the people of California.

January-2019 has the most with a count of almost greater than 15000 non-fatal collisions. We can clearly see from the graph above that the number of non-fatal collisions has reduced drastically as the pandemic got worse. The months from April 2019 till July 2020 had the least number of non-fatal accidents in the State of California. The number of non-fatal collisions started increasing gradually once people started to make movement across California. From the two graphs above we infer that the total number of fatal collisions are way more than the total number of non-fatal collisions. The non-fatal collisions include minor to mid-range of injuries.

V. CONCLUSION

Despite small number of vehicles operating in the year 2020, the level of crash accident recorded in California, made the state one of the top in the United States for traffic collisions. Through this data analysis a variety of insights concerning the location, time, weather, and points-of-interest of an accident are found. The analysis helps us understand the best month, day, and hour of the day to commute. Also, it can help us to predict what are the accident prone areas in the state such as Los Angeles, Kern and Orange with Highway and Highway road Intersections. It also shows that the highest death is happening between the 20- 50 age group and most of the accidents have occurred during a cloudy weather. The top 3 violations causing maximum collisions were: not following Traffic guidelines (11%), Unsafe speed (22% approx.) and Improper turns (17%).

REFERENCES

[1]E. Fitzgerald and B. Landfeldt, "Increasing road traffic throughput through dynamic traffic accident risk.

[2]Alyssa Ditcharoen, Bunna,Chhour, Tunyarat Traikunwaranon, NalinAphivongpanya, KunanonManeerat, Veeris Ammarapala, , "Road traffic accidents severity factors: A review paper" , 2018 5th International Conference on Business and Industrial Research (ICBIR).

[3]D. Mohan, G. Tiwari, and S. Mukherjee, "Urban traffic safety assessment: a case study of six Indian cities," IATSS Research, vol. 39, no. 2, pp. 95–101, 2016.

[4] Peden, M., et al. (2004) World Report on Road Traffic Injury Prevention. World Health Organization.