3-2020

ANALYSIS OF US ACCIDENTS AND SOLUTIONS

Swapnil Kisan Nikam
California State University - San Bernardino

ANALYSIS OF US ACCIDENTS AND SOLUTIONS

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Information Systems and Technology

---

by

Swapnil Kisan Nikam

March 2020

ANALYSIS OF US ACCIDENTS AND SOLUTIONS

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Swapnil Kisan Nikam

March 2020

Approved by:

Harold Dyck, Committee Chair, Information & Decision Sciences

Conrad Shayo, Committee Member

Jay Varzandeh, Committee Member

ABSTRACT

Reducing traffic accidents is an essential public safety challenge all over the world; therefore, accident analysis has been a subject of much research in recent decades.

The objective of the project is to analyze the US accident data from 50 states to inform the US government agencies and the general public on trends and possible causes of traffic accidents and what could be done to reduce them. The analysis include number of accidents by year, number of accidents by state, best time to travel by month, day and hour, accident-prone area in each state, factors responsible of the accidents like weather, wind flow, temperature, location, etc., deaths in each state, age group of fatalities, drivers involved in accident, drivers age group, vehicles involved in accident, driver with alcohol consumption.

The analysis platform is built using Tableau. Government agencies and the general public can leverage these insights and take a preventive measure which can reduce US accidents.

The significant findings from the analysis are: (a) most of the accidents are happening in October, November, December; (b) nighttime is safe to travel, (c) a more substantial number of deaths are from drivers in the 20-35 age group; (d) weather, temperature, and location are the factors responsible for 9% of accidents, and (e) about 60% accidents are attributed to drunk driving.

The main recommendations from the project focus on Infrastructure, Policy, Administrative, and Human behavior-related changes that can be implemented by the state and the federal government.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

After World War II, the automobile engine picked up the preeminent position as the primary means of transport. From that point forward, no one has challenged the dominance of engine vehicles. Instead, there have been various efforts to improve them, for example: to make the assembly line faster, to make them progressively adapted to the geographical terrain found in individual nations.

At present, automobile transport has become a piece of daily life. Improvement of automobiles is inescapable given the shockingly on-going high rates of terrible accidents and deaths. Unfortunately, vehicle crashes have always been a part of the vehicle driving experience. In 1771, Nicolas Joseph Cugnot caused the first road accident by crashing a self-built, improved version of the world's first steam-powered vehicle into a wall. The car was severely damaged as a result of the accident (Xu, Si-ji, Yan-ping, &Ye-Jiang,2005, p.2).

The first road casualty was an Anglican pastor from the town of Redruth. He died of fright at the sight of a loud and fast-moving model of "a steam engine on wheels," which was designed by William Murdock. This event took place in 1786 (Kiess, Rybicki, &Mauve, 2007, p.3). The world's first fatality caused by a road accident was reported to happen on August 31, 1869, in County Offaly in the UK(Goniewicz1, Goniewicz2, Pawłowski1, Fiedor,2015,p.1) While traveling as a passenger of an experimental steam-powered vehicle built by her cousins, an Irish Mary Ward fell out of the car on a bend, under its steel wheels and was run over. She died as a result of injuries (Anderson &Anderson,2004). The first USA road accident involving a vehicle fitted with an internal combustion engine took place in 1891 in Ohio City.

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. To show the significance of traffic accidents globally, the World Health Organization (WHO), in its worldwide status report on road safety 2009, estimates that in high-income nations like the USA there are 65 % of reported vehicle deaths from the Vehicle Occupants as compared to middle-income countries of the western pacific locale where 70% of the deaths are among vulnerable street users (WHO,2009). The same report also predicts that road traffic injuries will rise to become the 5th leading cause of death by 2030 (WHO,2009).

Although the global loss and suffering resulting from road accidents are indeed small compared with that caused by poverty and sickness, the problem is more severe than the present figures alone indicate. It is necessary to consider the monetary loss to nation-states due to fatal automobile accidents. A large number of the fatalities happen indiscriminately to vehicle users. In 2010, the economic loss of the USA alone was about 836 billion. (U.S. Department of Transportation,2015).These include educated individuals; the statesmen, specialists, instructors, and business people whose loss to the nation is severe.

PROBLEM STATEMENT

The present absence of extensive data systems equipped for gathering, classifying and detailing the accident and non-crash related injury information seriously confines the capacity to create, test, and implement alleviation strategies. The errand of recognizing injury causative components gets awfully theoretical without timely, accurate, complete, integrated, and available information that incorporates area, cause, contributing elements, and related activities associated with injuries involving personnel.

Government officials and general public are lacking systems which can show

1. What is the accident-prone area in each state?
2. What day and time are safe to travel?
3. What are the factors responsible for accidents?
4. What is the severity of these accidents?
5. How many deaths happening in accidents?
6. What solution can be implemented to reduce accidents by each state?
7. How can this accident be minimized?
8. How can the State Government improve accident-prone infrastructure?

CHAPTER TWO

LITERATURE REVIEW

There is a great deal of research out there that addresses accidents occurring the world over (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019). Regardless of all these progressing research numbers of accidents happening are not decreasing, which is a primary worry to everybody. However, the vast majority of them are on accident analysis, and prediction has utilized restricted assets that are not giving a full idea of the issue and affecting it to result which we need. In one of the research papers “A Countrywide Traffic Accident Dataset” (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019), they have tried to address this issue by collecting the data from API resources available from various sources and having records of 2.25 million instances of traffic accidents that took place within the contiguous

The United States, and over the last three years. Each accident record consists of a variety of intrinsic and contextual factors such as location, time, natural language description, weather, period-of-day, and points-of-interest (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019). Chang et al. (Chang, 2005) utilized information such as road geometry, annual average daily traffic, and weather data to predict the occurrences of accidents for a highway road by designing a neural network model.

Over time numerous studies have used large scale datasets; however, the datasets have been either private or not easily accessible (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019). Eisenberg (Eisenberg, 2004) carried analysis to identify the impact of road accidents with a large dataset of about 456000 crashes in 48 US states from 1975 to 2000. Recent studies by Najjar et al. (Najjar, Kaneko, & Miyanaga, 2017) have used large scale datasets to analyze real-time traffic accident prediction. Despite all these studies, results were not available for further research. The main thing about the dataset is, it is available publicly; however, it is limited in terms of one city or state, attributes are not enough for analysis.

Most of the research is not readily available to Government agencies and the public. If we consider all these researches, we can find that there is a big gap between the result found from this research and the implementation of this outcome.

To address these challenge, we propose a new platform which can showcase all the finding by each state like day and time safe to travel, accident-prone area and zip code in each state, severity, weather conditions, also if someone wants to go from Los Angeles to San Francisco in which area accidents mostly occur. For State Government officials, this platform will help to make a decision and provide solution-based on accident issues face by each state.

CHAPTER THREE

THE METHODOLOGY

This section presents a brief explanation of the data used and the methodology followed to accomplish results.

To address the problem question, the overall study is divided into three parts. First, understanding the current accident situation in the United States, like how many accidents are happening each year, deaths in accidents, the severity of accidents, the time and day safe to travel, and overview.

The second part will mainly focus on these accident issues concerning each state, like what are the accident-prone areas and the zip code in each state.

The third part will address the solution for the public as well as State Governments, how these accidents can be reduced, and what are solutions can be implemented by each state based on the crucial factors. This will be achieved by collecting data from various sources like the National Highway Traffic Safety Administration, State Government Websites, and Kaggle.com. After that, we can analyze the data and propose a possible solution to the problem question.

Data Sources And Collection

The data used in this research was obtained from various sources and covered the period between 1994 and 2019. Characteristics of traffic accidents were obtained from analyzing the available data. The data was collected, clean, manipulated, tabulated, and then analyzed. Most features were obtained for the period between 1994 and 2019, including the age of drivers, the severity of accidents, speed limit, a monthly distribution of accidents, accidents' distribution per governorate, among others.

Data Set Description

Kaggle.com has collected streaming traffic data from two sources, "MapQuest Traffic" (MapQuest Traffic API, 2019) and "Microsoft Bing Map Traffic" (Bing Map Traffic API, 2019) respectively, "whose APIs broadcast traffic events (accident, congestion, etc.) captured by a variety of entities - the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks" (Moosavi, Samavatian, Nandi, Parthasarathy, &Rajiv Ramnath,2019). There are 3 million records in this dataset from February 2016 to December 2019.

Table 4-1:Data Set Description(US-Accidents: A Countrywide Traffic Accident Dataset,2019)

#	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).

3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.
4	Severity	Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).
5	Start_Time	Shows start time of the accident in local time zone.
6	End_Time	Shows end time of the accident in local time zone.
7	Start_Lat	Shows latitude in GPS coordinate of the start point.
8	Start_Lng	Shows longitude in GPS coordinate of the start point.
9	End_Lat	Shows latitude in GPS coordinate of the end point.
10	End_Lng	Shows longitude in GPS coordinate of the end point.
11	Distance(mi)	The length of the road extent affected by the accident.

12	Description	Shows natural language description of the accident.
13	Number	Shows the street number in address field.
14	Street	Shows the street name in address field.
15	Side	Shows the relative side of the street (Right/Left) in address field.
16	City	Shows the city in address field.
17	County	Shows the county in address field.
18	State	Shows the state in address field.
19	Zipcode	Shows the zipcode in address field.
20	Country	Shows the country in address field.
21	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).
24	Temperature(F)	Shows the temperature (in Fahrenheit).
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
26	Humidity(%)	Shows the humidity (in percentage).
27	Pressure(in)	Shows the air pressure (in inches).

28	Visibility(mi)	Shows visibility (in miles).
29	Wind_Direction	Shows wind direction.
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
33	Amenity	A POI annotation which indicates presence of amenity in a nearby location.
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
35	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
36	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
37	Junction	A POI annotation which indicates presence of junction in a nearby location.
38	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.

39	Railway	A POI annotation which indicates presence of railway in a nearby location.
40	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
41	Station	A POI annotation which indicates presence of station in a nearby location.
42	Stop	A POI annotation which indicates presence of stop in a nearby location.
43	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
44	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.
45	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
46	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
47	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
48	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.

49	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.

Data related to the number of deaths in accidents from 2004 to 2018 has extracted from the National Highway Traffic Safety Administration site along with deaths by age groups and drivers with alcohol who involved in accidents taken to analyze factors for accidents more appropriately.

Tools

Python

Python is an easy to understand and common use programming language. Python is an object-oriented language used for data analysis.

Benefits of Python:

- Data Analysis
- Development of Website
- Development of Application

Why Python?

- Python is available on different platforms (Windows, Mac, Linux, Raspberry Pi, etc.).

- Python is more straightforward as the English language, i.e., simple coding.
- Python has a syntax that allows developers to write programs with fewer lines than some other programming languages.
- Python runs on an interpreter system, meaning that code can be executed as soon as it is written. This means that prototyping can be rapid.
- Python can be treated procedurally, an object-orientated way, or a practical way.

In this project, python has used for data cleaning and data manipulation. Jupyter Notebook has used for python. All analysis is done in Python 3.0.

Tableau

Tableau is heavily used in business intelligence. Users can make and disperse an intelligent and shareable dashboard, which delineate the patterns, varieties, and thickness of the information as graphs and charts. Tableau can associate with files, relational, and Big Data sources to get and process information. The software permits information blending and real-time collaboration, which makes it extremely special. Organizations utilize it, academic researchers, and numerous administration associations for visual information examination Tableau can create a data visualization, data analytics, and reporting by just dragging and dropping columns. Tableau does not require an earlier coding experience. Tableau can import data from various data sources like databases, spreadsheets, big data, and cloud data into one program to perform dynamic analysis.

Why Tableau?

Whether it is small or large, profitable, or non-profit, every organization needs to analyze their data for optimal decision making. Analyzing data has never been more comfortable with traditional business intelligence tools.

Table 4-2:Comparison of Traditional Method and Tableau (Rahman, 2015)

Traditional Method	Tableau
Prior programming skills	No programming skills required
Focused on only one type of database	Combines different types of database spreadsheets, databases, cloud data, and even big data such as Hadoop
Decision-makers have to ask the IT people to retrieve any information from the database	Decision-makers can directly use the dashboard to retrieve any information from the database
Mostly depends on Query languages	The query is done behind the scene
Combining different types of the database is difficult	Different types of databases can be combined easily

Not every business intelligence tool offers an interactive dashboard	The interactive dashboard is easy to build, and it makes data visualization quick and efficient
Mostly designed for large businesses	Perfect BI solution for small, medium, and large businesses, and even for non-profits
Comparatively expensive	Comparatively affordable
Time-consuming	Time-saving

Data Cleaning

While the data set has 3 million records, it is not ready to use for analysis. There are many anomalies in the dataset like:

- Null records
- Date format
- Day is missing
- Duplicate records
- Mismatched column

To Address all these anomalies in data, data cleaning is the most important and mandatory step. Data cleaning is performed as below in the analysis:

- Null records

  Considering the percentage of null records depending on the category. Some of the null records have been deleted from the dataset and been replaced with appropriate values.

- Date format

  The date format was not consistent, as data was collected from many resources. For the analysis, it is necessary to have data in a logical form, so all date formats were changed to “MM/DD/YYYY.” To clean the date column, the calculated field was created in Tableau.

- Day is missing

  In the dataset day of the accident, the day was not mentioned. To understand the trend of accidents over the week, it is essential to know on which day the accident has happened. So a calculated field was created in Tableau.

- Duplicate records

  Dataset had many duplicate records, and all duplicate records were deleted to gain more accuracy.

- Mismatch column

  When data is imported in Tableau, it automatically identifies the data type of the column; however, in our dataset tableau, it was unable to locate some of the columns. As part of the data cleaning process, these column datatypes were corrected.

CHAPTER FOUR

DATA ANALYSIS AND VISUALIZATION

US Accidents At A Glance

<u>Number of Accidents</u>

Number of Accidents were lower and showed a more significant change in the early years as compared with those in recent years when the number of accidents has been raised high.

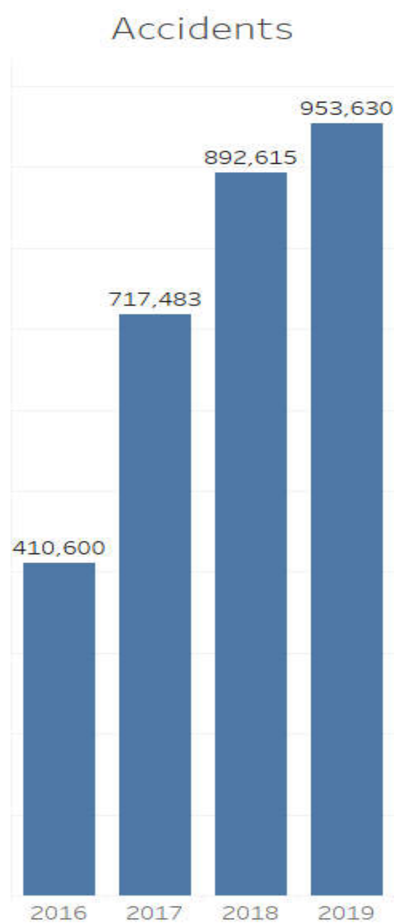

Figure 5-1:Accident Trend Over the Years

Number of Accidents by State

Figure 5-2 showing the number of accidents by each state. The color scheme used to showcase a difference between the number of accidents that happened in each state over the years.

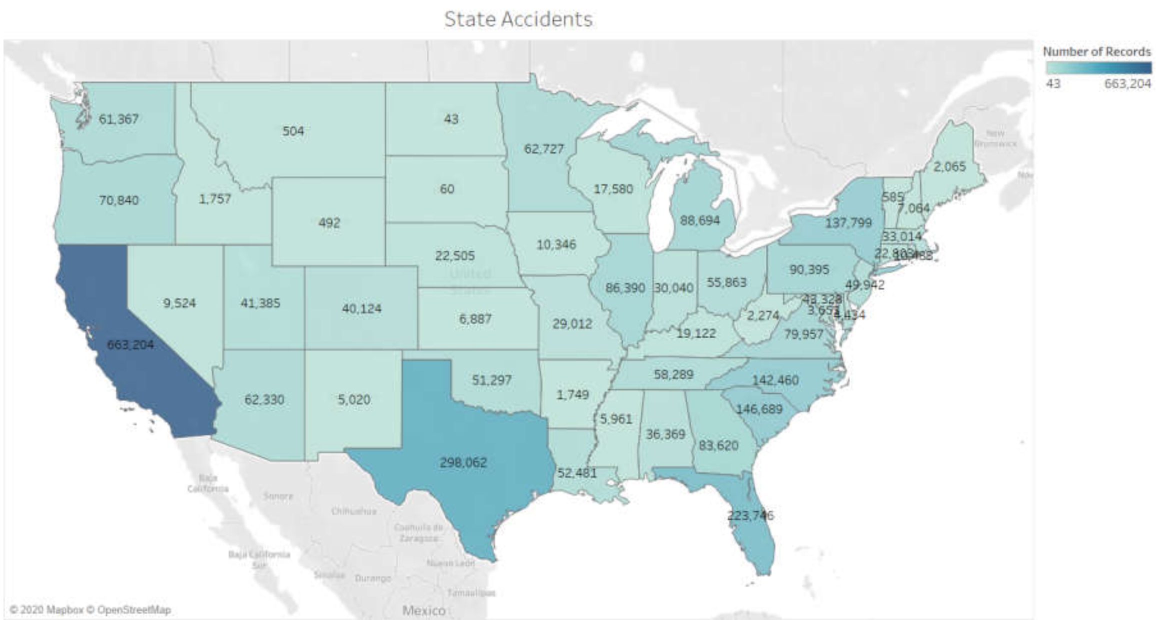

Figure 5-2:Number of Accidents in Each State

Top five states in which most of the accidents happening were showing below figure 5-3.

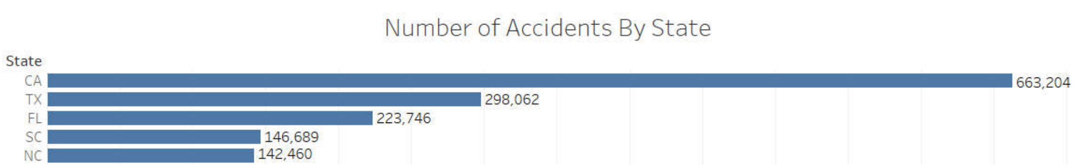

Figure 5-3:Top Five States with a High Number of Accidents

Accidents By Zipcode

Figure 5-4 showing visualization of a number of accidents happen by zip code. East and west coast has more accident compare to central areas. Texas is also showing a large number of accidents.

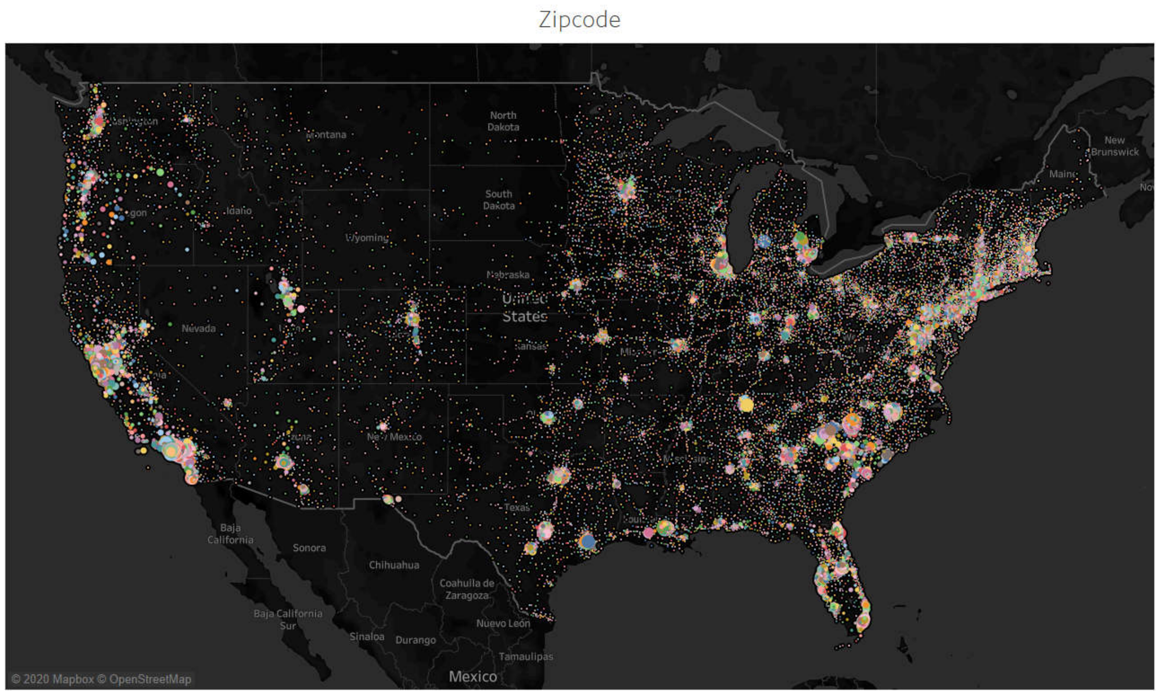

Figure 5-4:Accidents by Zipcode

Below are the top ten zip codes with their state and count of accidents happen. North Carolina has five zip codes in the top 10 lists, along with California, South Carolina, Texas, Michigan, and Louisiana.

Top Accident Prone Zipcode with State

State (U..	CleanZip..	
CA	91706	4,854
LA	70808	4,217
MI	48507	4,246
NC	28208	5,677
	27610	5,592
	28205	5,117
	28216	4,975
	27604	4,577
SC	29210	4,276
TX	78753	4,388

Number of Records
4,217 5,677

Figure 5-5:Top Ten Accident Prone Zipcode with State

<u>The Trend of Accidents by Month</u>

Monthly accidents increase steadily from the lowest points in January and February, peak in October-December. First half showing a smaller number of accidents compared to the second half of the year. Monthly accident rates steadily increase from the lowest points in February, and peak in the last quarter of the year.

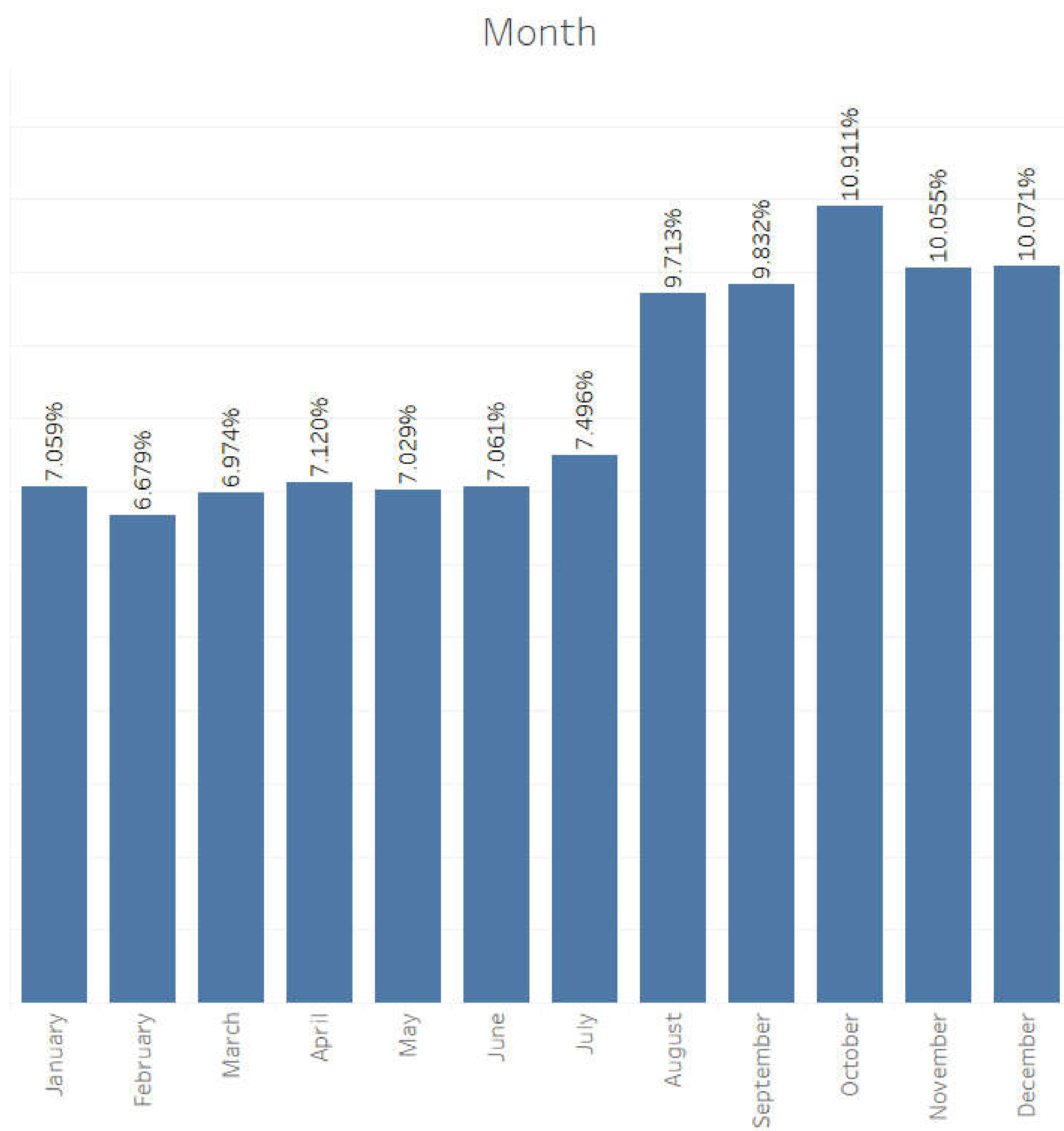

Figure 5-6:Trend of Accidents by Month

The Trend of Accidents by Day

Between 2016 and 2019, there is a large number of accidents during weekdays (Monday to Thursday). On the contrary, there are relatively fewer accidents on weekends (Friday, Saturday, and Sunday)

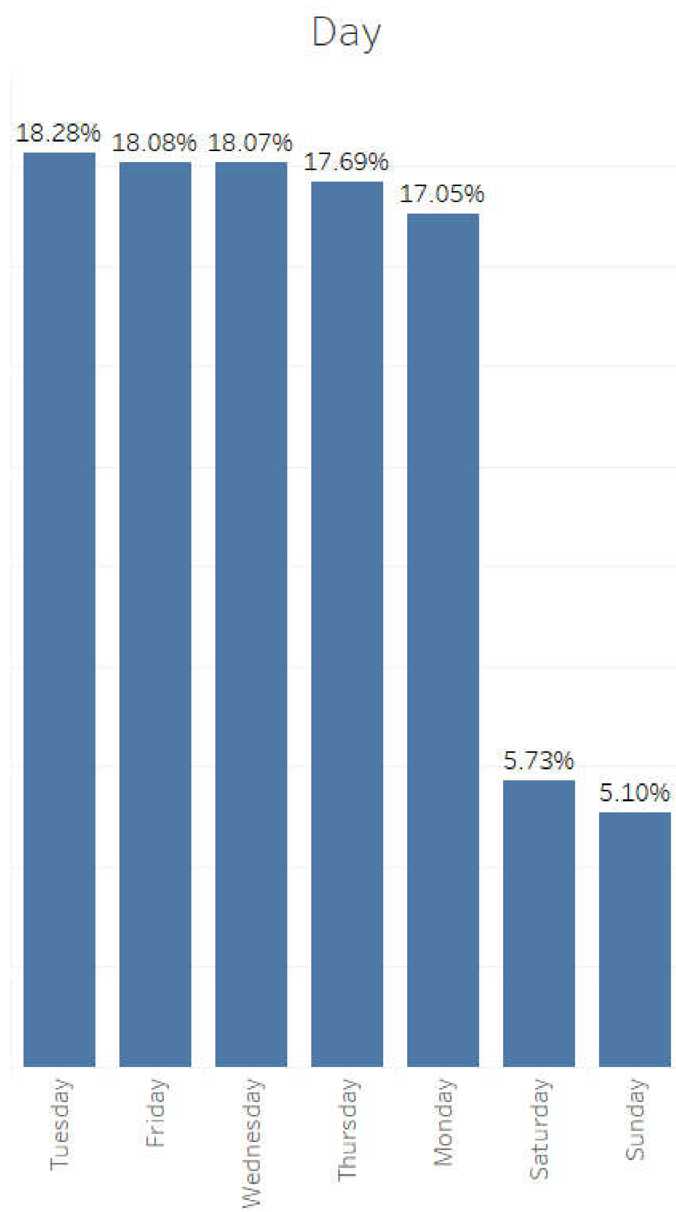

Figure 5-7:Trend of Accidents by Day

The Trend of Accidents by Hour

Between 2016 and 2019, peak hours of the accidents are 7 AM, 8 AM, 5 PM, and 6 PM and nighttime (10 PM-5 AM) is the safest time to travel which is showing less number of accidents.

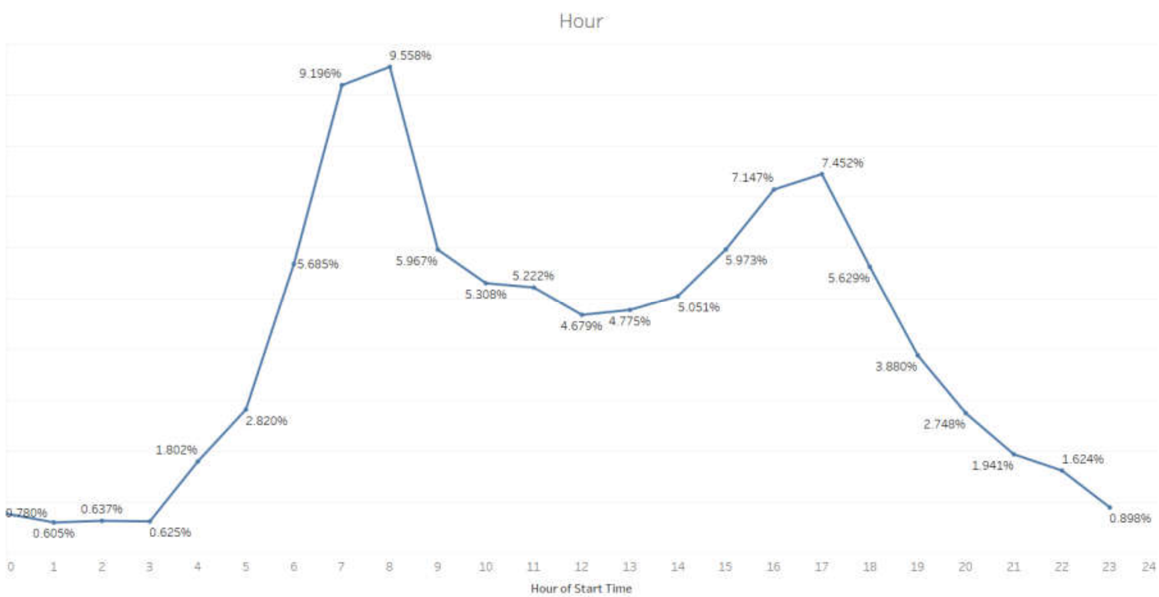

Figure 5-8:Accidents By the time of Day

Accident Factors

Accident Severity

Severity indicates the impact of the accident. Severity 3&4 having a high impact. Severity 2 accidents are more compare to severity 3&4. Severity 3&4 accidents mostly involve fatalities and severity 2, having injuries. Severity 1&2 accidents are reported primarily due to insurance claims.

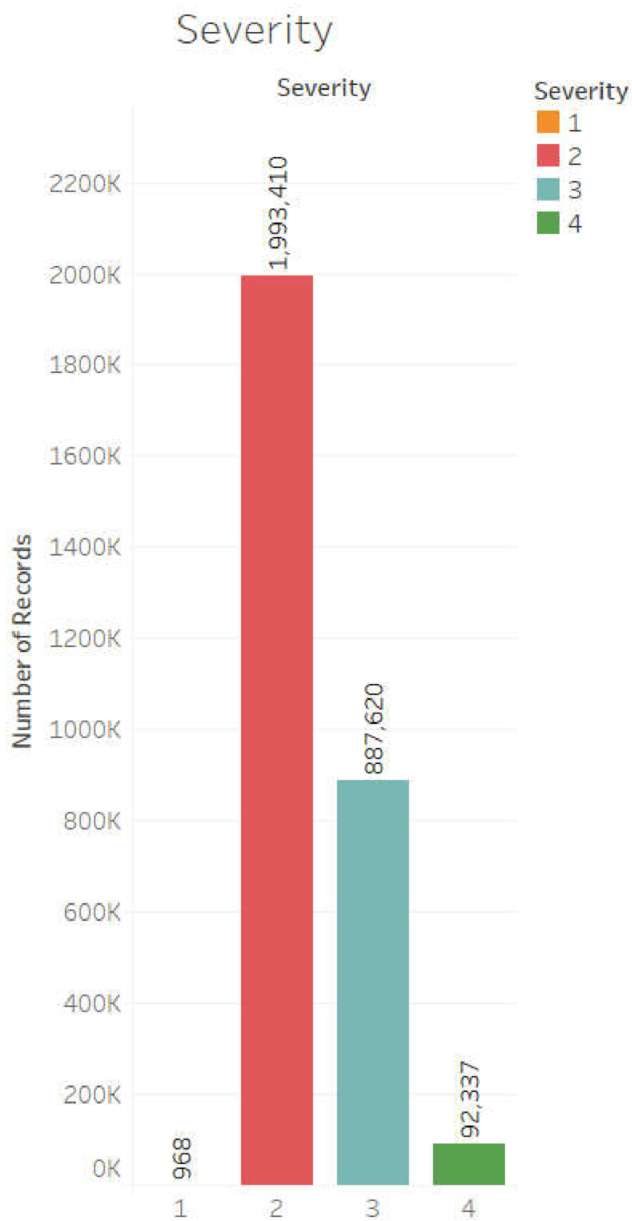

Figure 5-9:Accidents by Severity

Accidents by Timezone

Figure 5-10 shows the percentage of accidents in each timezone. Eastern time zone having the highest number of accidents around 43% than pacific, central, and Mountain time zone.

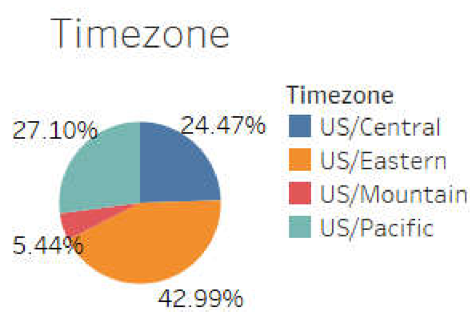

Figure 5-10:Accident Percentage by Time Zone

Accidents by Temp

Temp is not a crucial factor in deciding the accident. Figure showing histogram of accident temp.

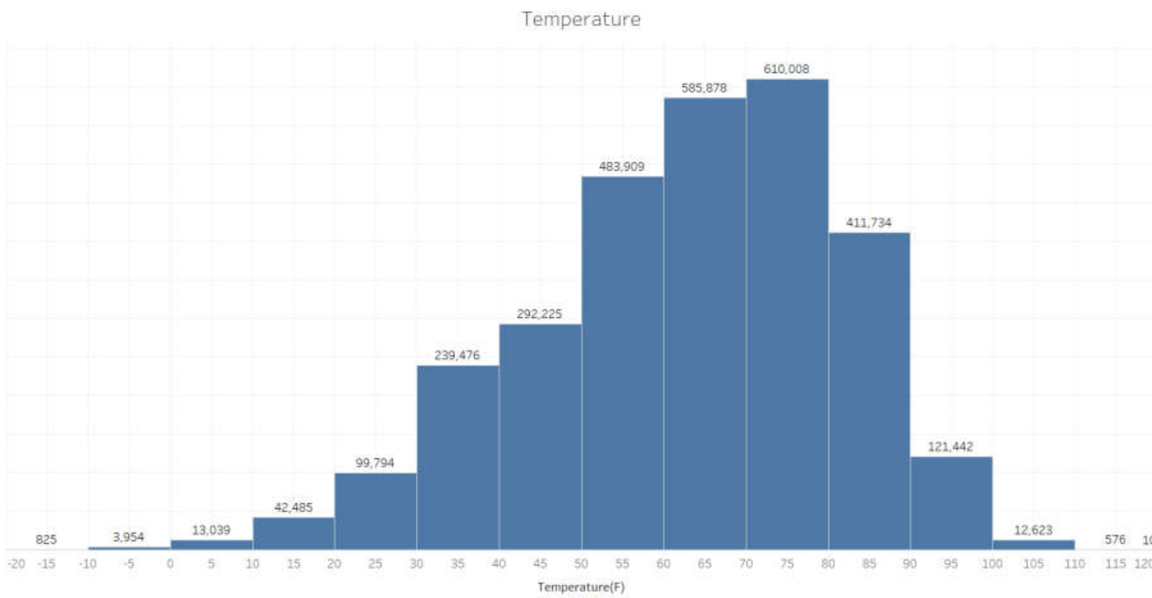

Figure 5-11:Temp at the Time of Accident

Accident Correlation

Figure 5-12 showing the accident correlation matrix between various factors. From the matrix, for accident Start Lat and End Lat are the crucial factors for the accident. Temp and wind flow are playing an important role.

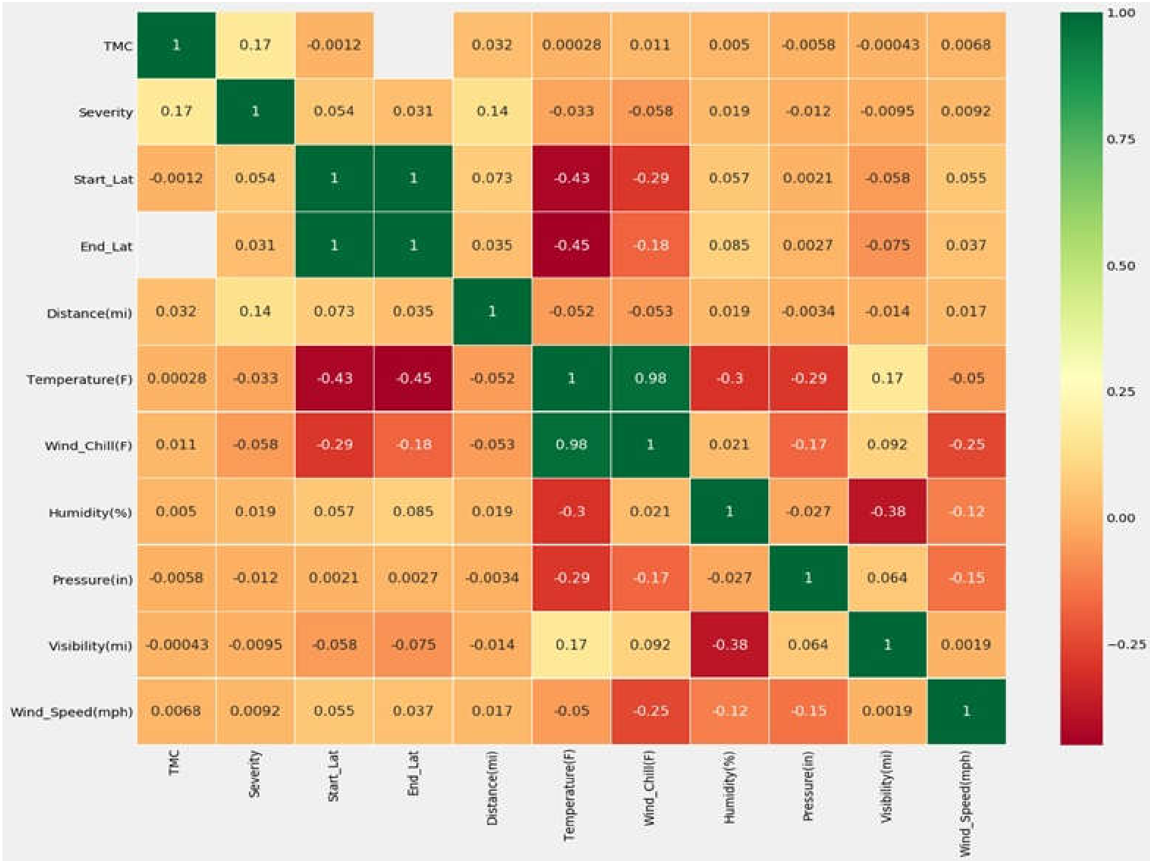

Figure 5-12:Accident Correlation Matrix

Accident Description Word Cloud

Word Clouds are visual representations of words that give greater prominence to words that appear more frequently. Word cloud helps presenters to quickly collect data from their audience, highlight the most common answers, and present the data in a way that everyone can understand.

Based on accident description data, the word cloud has plotted and exit, southbound, northbound, westbound, eastbound, highway, and parkway are the most common words which indicate that most of the accidents happening there.

Figure 5-13:Accident Description Word Cloud

Accident Stats by State

<u>Deaths in Accident</u>

Figure 5-14 showing the fatality trend in accidents from 2004 to 2018. In the early years, it was high; after that, it was lower in the middle year and gained it started to rise from 2016.

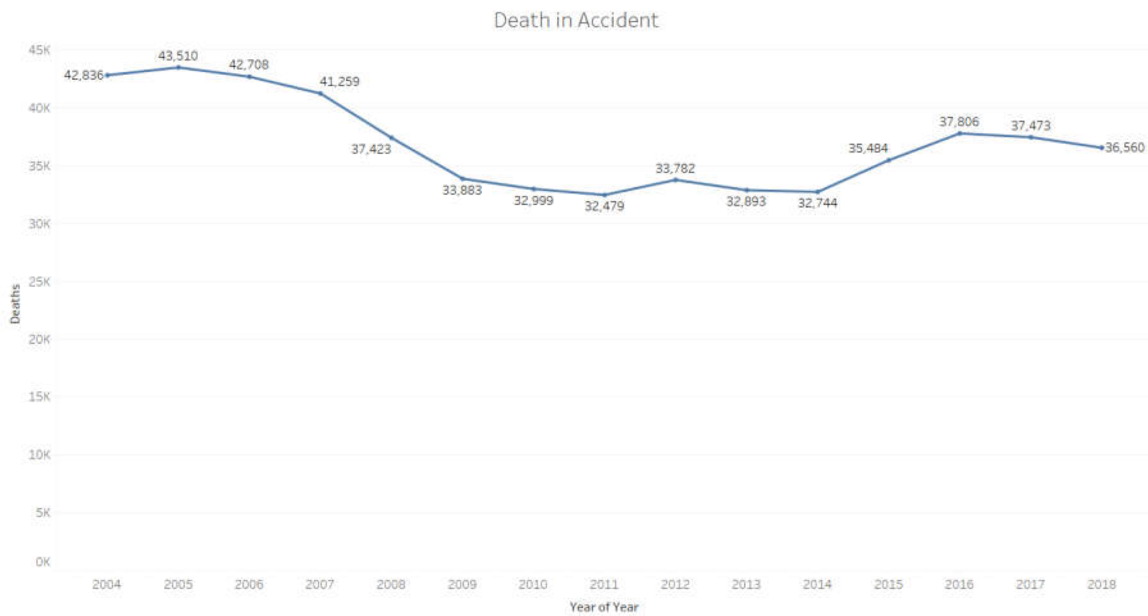

Figure 5-14:Fatality in Accident (2004-2018)

<u>Deaths in Accidents by State</u>

The highest number of deaths happened in California and then in Texas compare to other states.

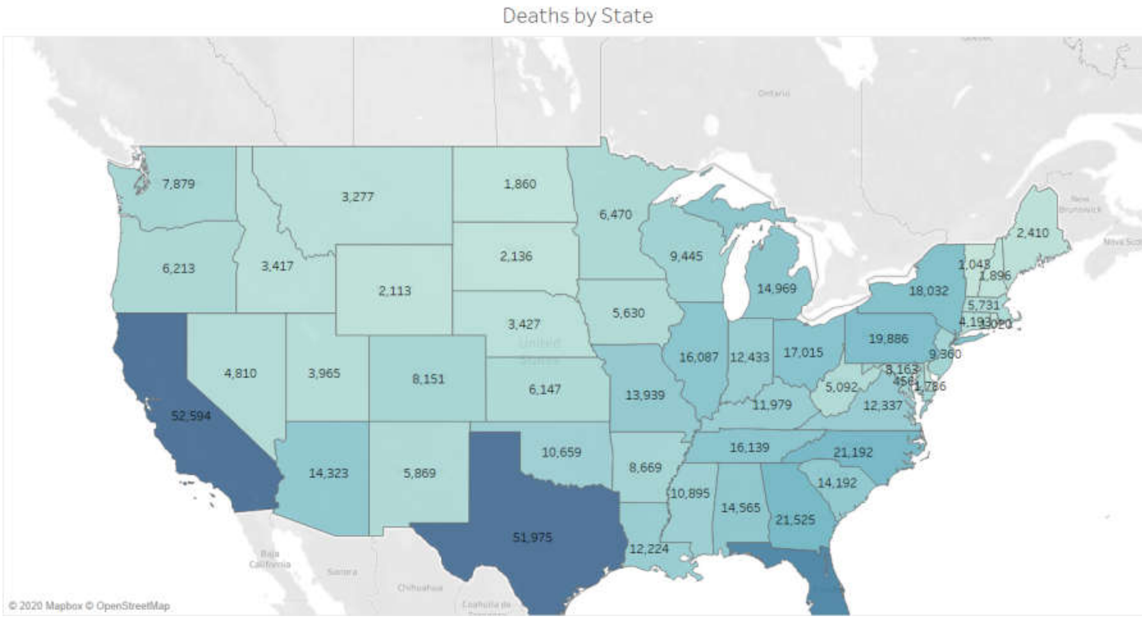

Figure 5-15:Fatality in Accident by State (2004-2018)

Deaths by Age Group

Figure 5-16 showing deaths by various age groups. Most of the deaths happened in the 25-34 age group. This age group is contributing to the economy of the country, and which is a topic of concern.

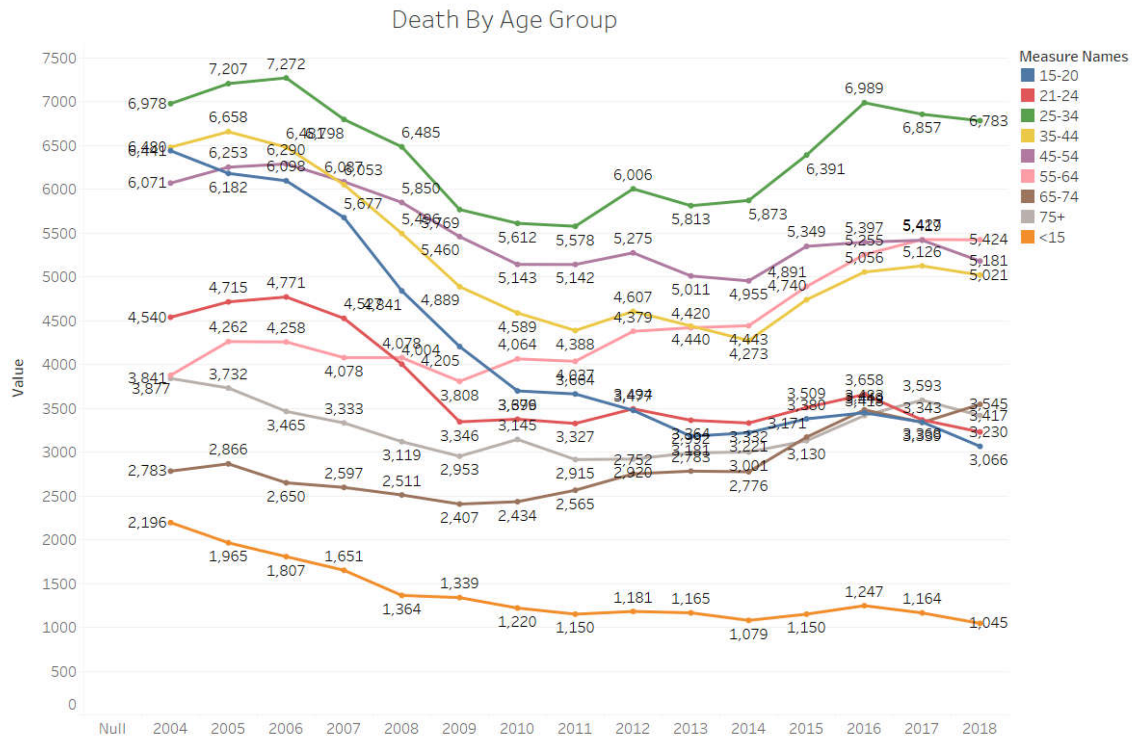

Figure 5-16:Fatality by Age Group in Accident (2004-2018)

Drivers Age Group

Driver Age group

Year of Crash ..	<15	15-20	21-24	25-34	35-44	45-54	55-64	65-74	75+
2004	158	7,942	6,413	11,242	10,743	9,148	5,612	3,070	3,169
2005	138	7,500	6,585	11,467	10,793	9,434	6,075	3,217	3,016
2006	99	7,493	6,480	11,279	10,379	9,234	5,894	3,029	2,967
2007	107	7,026	6,287	10,773	9,936	9,028	6,037	3,038	2,879
2008	79	5,886	5,342	9,800	8,806	8,355	5,717	2,927	2,672
2009	84	5,170	4,612	8,630	7,779	7,686	5,276	2,876	2,560
2010	61	4,603	4,608	8,567	7,333	7,517	5,577	2,902	2,688
2011	60	4,362	4,488	8,549	7,084	7,513	5,572	2,960	2,528
2012	49	4,313	4,765	9,019	7,365	7,660	5,930	3,239	2,554
2013	56	3,991	4,630	8,808	7,220	7,376	5,947	3,373	2,586
2014	55	3,897	4,664	8,992	6,910	7,370	6,004	3,316	2,650
2015	61	4,352	5,014	9,994	7,768	7,915	6,525	3,794	2,762
2016	76	4,555	5,284	10,913	8,179	8,023	7,037	4,155	3,014
2017	62	4,410	5,070	11,006	8,284	8,186	7,316	4,148	3,151
2018	43	4,144	4,777	10,738	8,110	7,863	7,261	4,218	3,098
Grand Total	1,188	79,644	79,019	149,777	126,689	122,308	91,780	50,262	42,294

Figure 5-17:Driver Involved in Fatal Crash

	Driver Age group								
Year of Crash ..	<15	15-20	21-24	25-34	35-44	45-54	55-64	65-74	75+
2004	5,909	570,395	380,060	691,327	646,798	542,358	301,392	152,311	117,584
2005	12,464	513,528	374,159	682,796	620,845	535,475	292,303	141,956	107,995
2006	19,563	504,429	347,373	648,351	600,351	504,077	292,886	149,068	109,966
2007	10,180	470,766	346,251	644,590	547,203	489,980	303,747	145,427	99,113
2008	13,342	420,333	330,128	607,894	508,004	471,665	307,679	133,696	96,802
2009	15,675	372,732	319,220	556,237	475,853	460,372	287,507	138,724	97,855
2010	3,294	358,179	331,178	566,038	503,399	460,428	300,884	149,953	108,023
2011	3,134	346,958	298,399	585,288	490,140	448,695	331,138	150,824	95,886
2012	5,579	356,707	320,162	632,364	507,669	491,697	355,383	174,275	105,583
2013	5,918	355,157	338,465	624,436	488,465	463,154	352,799	178,416	108,279
2014	28,537	353,758	335,104	664,865	505,300	472,299	366,097	178,335	106,352
2015	30,138	381,036	353,908	684,156	534,994	487,328	385,911	201,564	111,710
2016	2,715	463,878	436,364	888,317	657,684	599,667	460,821	232,907	133,523
2017	3,031	418,437	377,930	779,965	594,960	534,533	426,452	235,134	125,962
2018	4,727	382,183	367,923	798,164	602,545	535,650	430,562	242,118	117,970
Grand Total	164,206	6,268,476	5,256,624	10,054,788	8,284,210	7,497,378	5,195,561	2,604,708	1,642,603

Figure 5-18:Driver Involved in Injury

	Driver Age group								
Year of Crash ..	<15	15-20	21-24	25-34	35-44	45-54	55-64	65-74	75+
2004	6,976	1,408,122	816,980	1,547,566	1,378,727	1,145,364	634,283	315,152	213,943
2005	114,222	1,177,745	811,026	1,555,342	1,379,057	1,255,647	653,075	324,173	221,375
2006	110,079	1,109,431	773,283	1,470,516	1,357,375	1,291,471	683,779	315,978	212,712
2007	11,653	1,152,824	769,966	1,520,538	1,311,850	1,262,034	848,980	311,367	221,707
2008	74,147	1,003,009	846,036	1,462,715	1,387,364	1,136,442	696,492	322,750	212,279
2009	119,598	959,197	822,446	1,343,012	1,210,521	1,135,156	701,315	338,709	214,779
2010	11,153	952,140	784,645	1,336,619	1,241,603	1,086,515	719,444	359,034	225,482
2011	2,640	877,870	747,279	1,350,151	1,237,005	1,044,511	768,836	351,833	215,552
2012	13,051	897,304	816,042	1,466,731	1,168,147	1,102,594	796,570	387,249	239,238
2013	10,609	913,511	800,632	1,519,957	1,167,349	1,169,662	843,108	422,938	235,090
2014	62,864	959,917	862,461	1,656,692	1,291,437	1,225,791	945,860	458,438	254,332
2015	79,698	1,020,274	901,858	1,740,910	1,354,017	1,198,726	973,688	499,798	263,121
2016	2,780	1,042,067	928,462	1,789,721	1,391,517	1,252,355	973,732	506,732	258,020
2017	8,839	1,005,224	876,132	1,779,143	1,350,323	1,220,142	956,081	526,208	251,549
2018	6,209	1,025,846	872,186	1,873,349	1,471,310	1,308,930	1,067,403	572,307	294,021
Grand Total	634,518	15,504,481	12,429,434	23,412,962	19,697,602	17,835,340	12,262,646	6,012,666	3,533,200

Figure 5-19:Drivers Involved in Property Damage

Driver with Alcohol Trend

Dataset has recorded with alcohol consumption as BAC=0.00, BAC=0.01-0.07, and BAC=0.08+. Drivers with BAC=0.01-0.07 and BAC=0.08+ considered as drunk while driving. Figure showing a number of such drivers from 1994-2018. The trend is uniform over the years, which is a concern for authority.

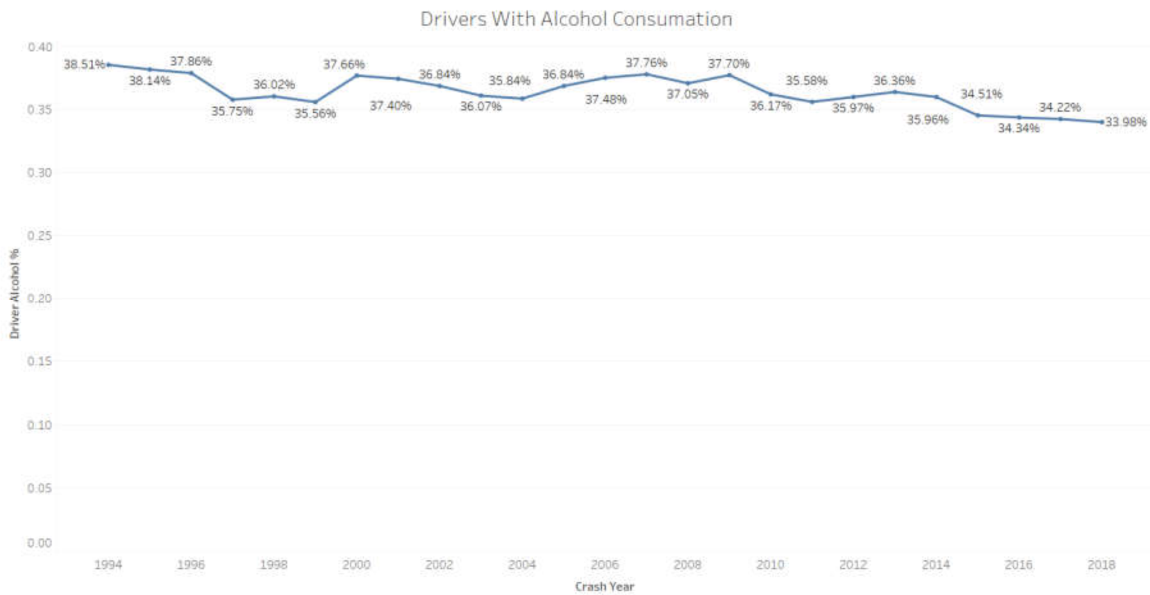

Figure 5-20:Alcoholic Drivers Involved in Accident (1994-2018)

Driver with Alcohol by State

Figure 5-21 showing the number of drivers found drunk when the accident happened. Texas has the highest number of cases, followed by California.

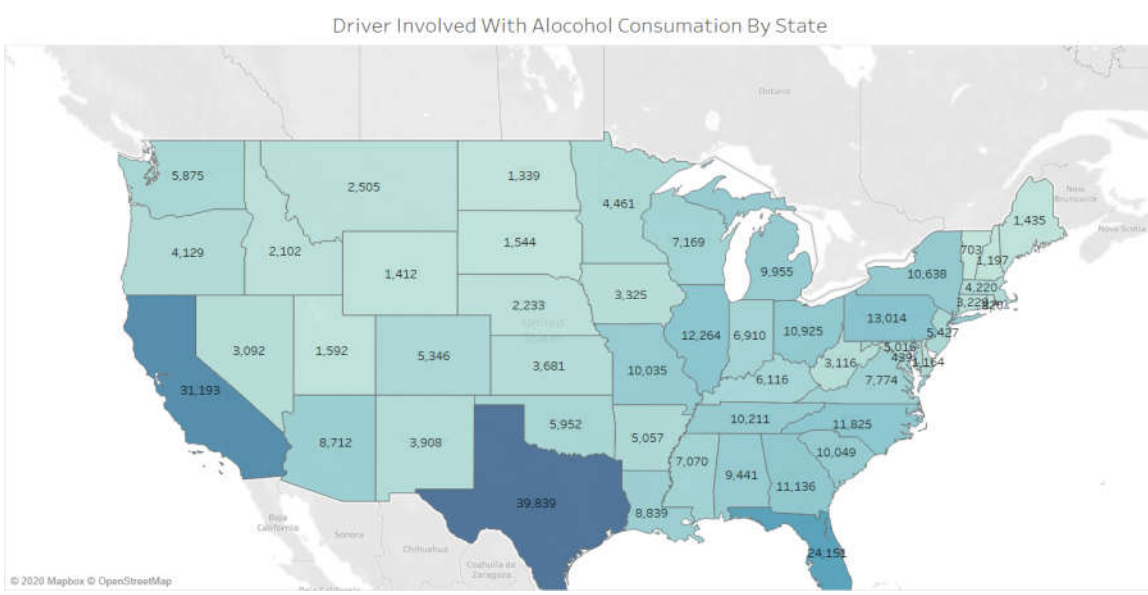

Figure 5-21:Alcoholic Drivers Involved in Accident By State (1994-2018)

Vehicles Involved in Accident Trend

Figure 5-22 showing the number of vehicles involved in the accident from 2004-2018. In the early years, it was high; after that, it was lower in the middle year and gained it started to rise from 2016.

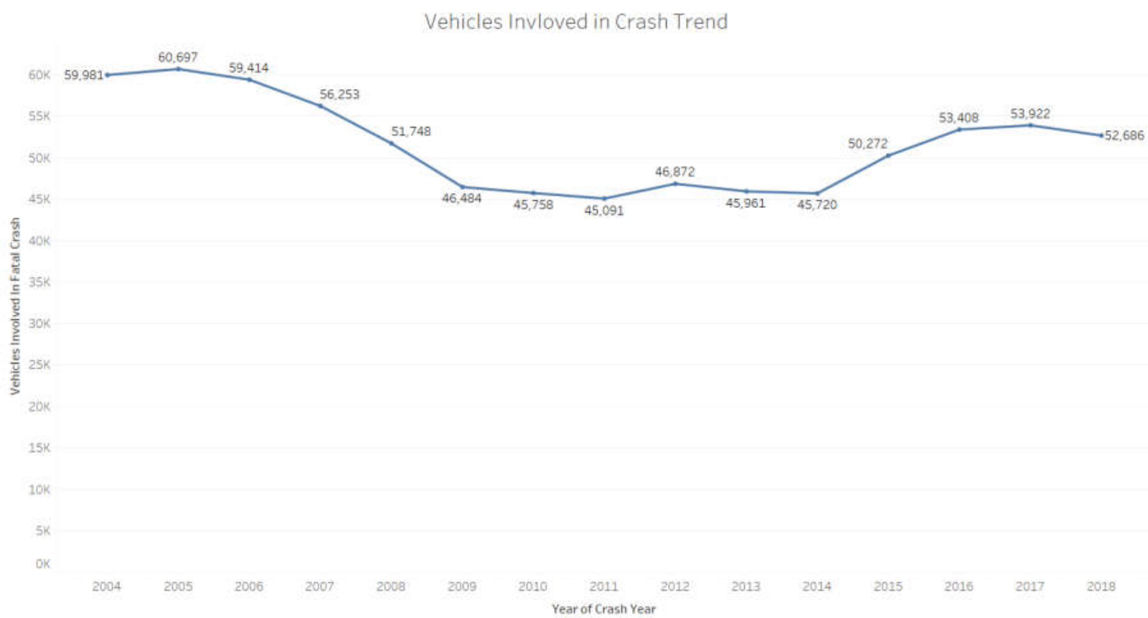

Figure 5-22:Vehicles Involved in Accident (2004-2018)

Vehicles Involved in Accident by State

Figure 5-23 showing the number of vehicles involved in an accident from 2004-2018 by state. California is leading in the list, followed by Texas.

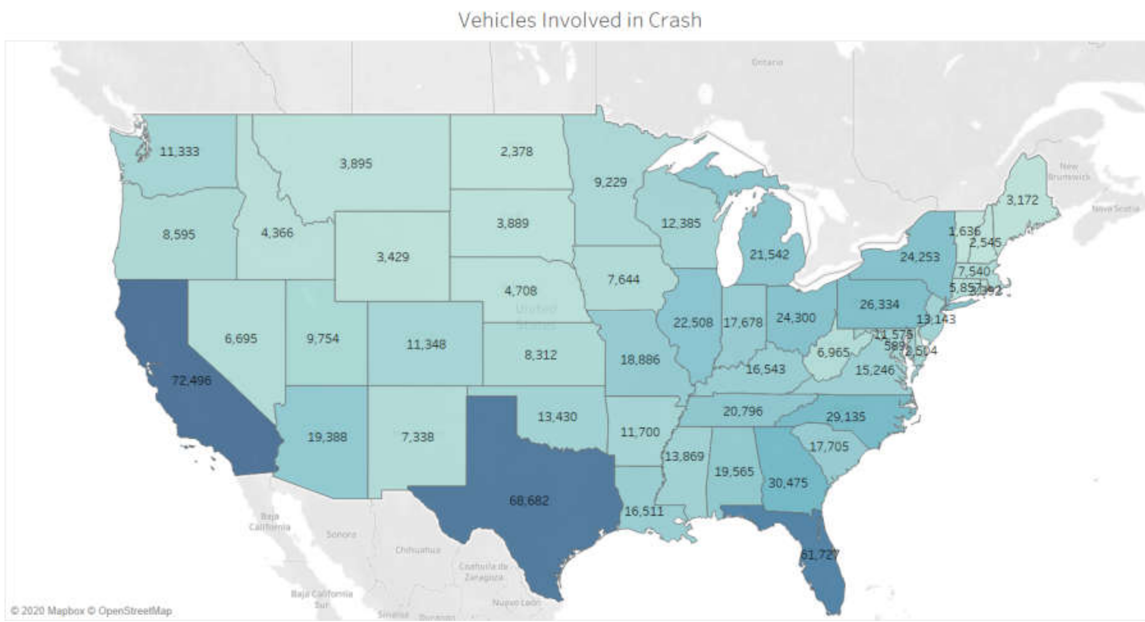

Figure 5-23:Vehicles Involved in Accident by State (2004-2018)

Forecasting

In this technique, future trends can be predicted based on historical data. Most of the organizations utilize forecasting to learn how to allocate their budgets or plan for anticipated expenses for an upcoming period.

<u>Death by Age Group Forecasting</u>

Using tableau forecasting has done to predict death by age for 2019 and 2020. As per forecasting for the age group <15 number is decreasing in the future; however, for the rest of the age group, it will remain the same.

Figure 5-24 showing forecasting as below

Death By Age Group

Year of C..	Actual									Estimate								
	<15	15-20	21-24	25-34	35-44	45-54	55-64	65-74	75+	<15	15-20	21-24	25-34	35-44	45-54	55-64	65-74	75+
Null																		
2004	2,196	6,441	4,540	6,978	6,480	6,071	3,877	2,783	3,841									
2005	1,965	6,182	4,715	7,207	6,658	6,253	4,262	2,866	3,732									
2006	1,807	6,098	4,771	7,272	6,481	6,290	4,258	2,650	3,465									
2007	1,651	5,677	4,527	6,798	6,053	6,087	4,078	2,597	3,333									
2008	1,364	4,841	4,004	6,485	5,496	5,850	4,078	2,511	3,119									
2009	1,339	4,205	3,346	5,769	4,889	5,460	3,808	2,407	2,953									
2010	1,220	3,699	3,376	5,612	4,589	5,143	4,064	2,434	3,145									
2011	1,150	3,664	3,327	5,578	4,388	5,142	4,037	2,565	2,915									
2012	1,181	3,477	3,494	6,006	4,607	5,275	4,379	2,752	2,920									
2013	1,165	3,181	3,364	5,813	4,440	5,011	4,420	2,783	2,992									
2014	1,079	3,221	3,332	5,873	4,273	4,955	4,443	2,776	3,001									
2015	1,150	3,380	3,509	6,391	4,740	5,349	4,891	3,171	3,130									
2016	1,247	3,449	3,658	6,989	5,056	5,397	5,255	3,483	3,418									
2017	1,164	3,343	3,368	6,857	5,126	5,419	5,427	3,339	3,593									
2018										1,085	3,118	3,327	6,707	5,337	5,359	5,761	3,279	3,416
2019										1,040	2,996	3,262	6,707	5,622	5,359	6,119	3,279	3,416
2020										996	2,879	3,199	6,707	5,922	5,359	6,498	3,279	3,416

Figure 5-24:Fatality By Age Group in Accident Forecasting (2019-2020)

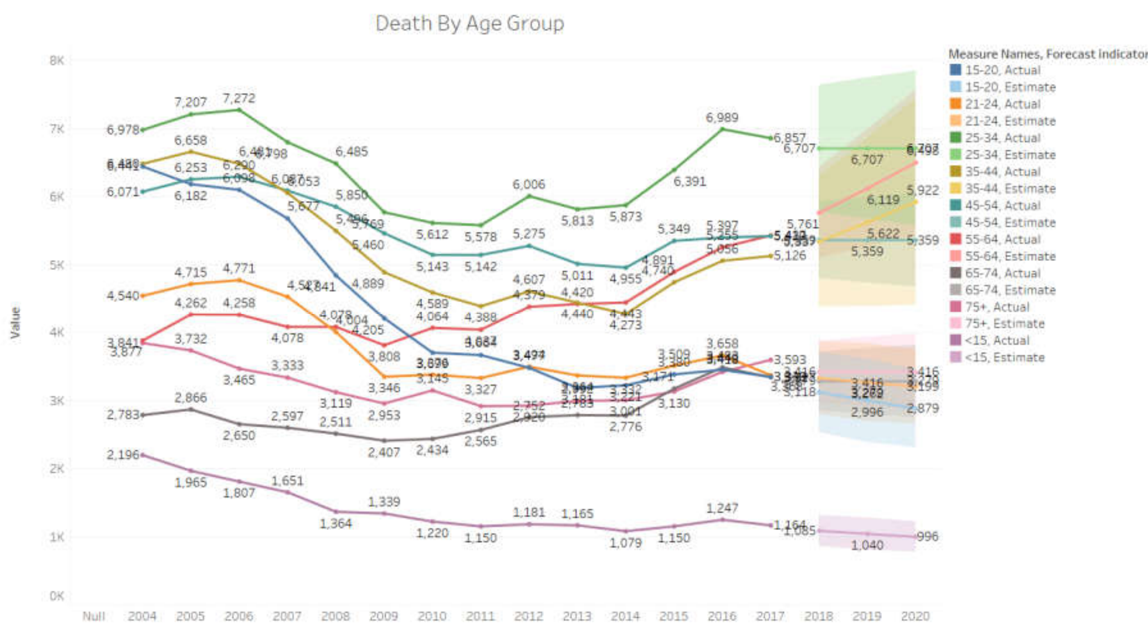

Figure 5-25:Fatality by Age Group in Accident Forecasting (2019-2020)

<u>Forecasting for Vehicles Involved in the Accident</u>

Figure 5-26 shows forecasting for Vehicles involved in the accident. It is showing the same for 2019 and 2020.

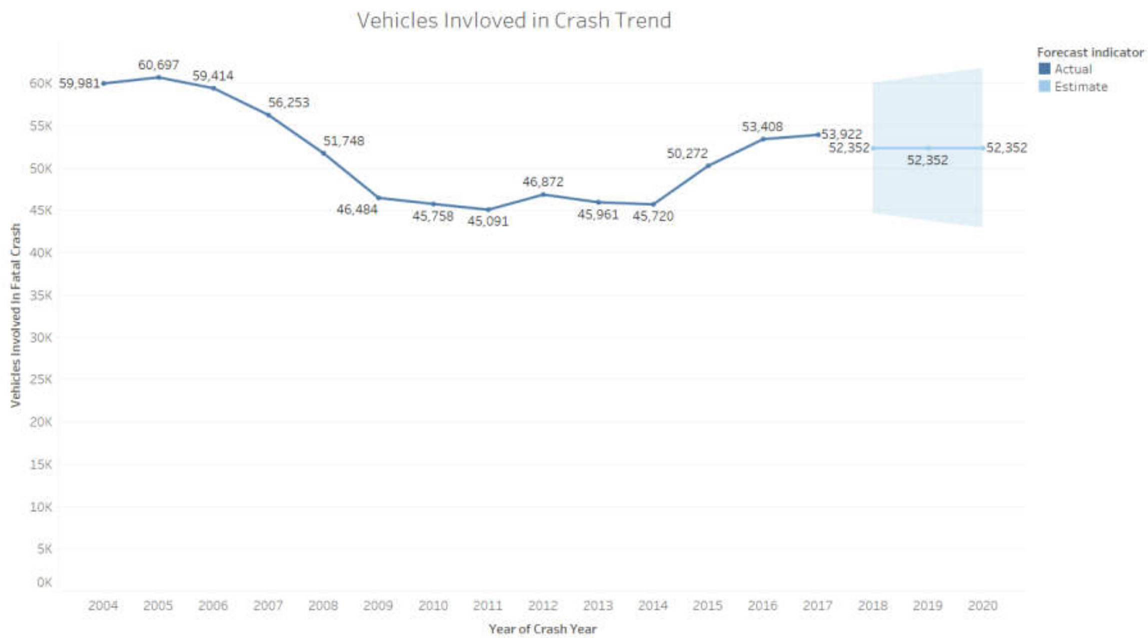

Figure 5-26:Vehicles Involved in Accident Forecasting (2019-2020)

Accident Stats Platform

Tableau based platform created, which can address questions related to US accidents. This platform is divided into four parts:

- US Accidents at a Glance
- Accidents Factors
- Accidents Stats By State
- The solution to an accident based on insights

The platform will be published on Tableau Public Profile so that it can be easily accessible. The platform is available at https://public.tableau.com/profile/swapnil.nikam4601#!/vizhome/USAccidentsAnalysisandSolution/Story1

- US Accidents at a Glance

  This dashboard will give an overview of the accidents over the years, like a number of accidents by each state, best month, day, and time to travel.

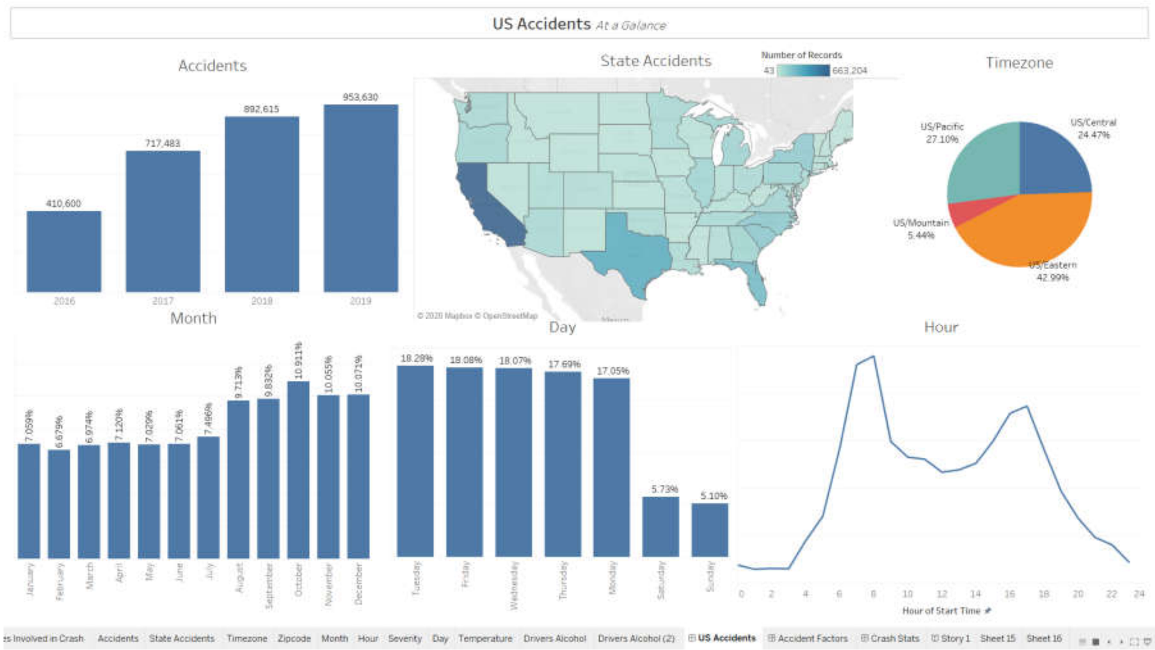

Figure 5-27:US Accidents at a Glance

- Accidents Factors

  This dashboard will focus on accident factors like severity, temperature, correlation, word cloud, and accident-prone zip code.

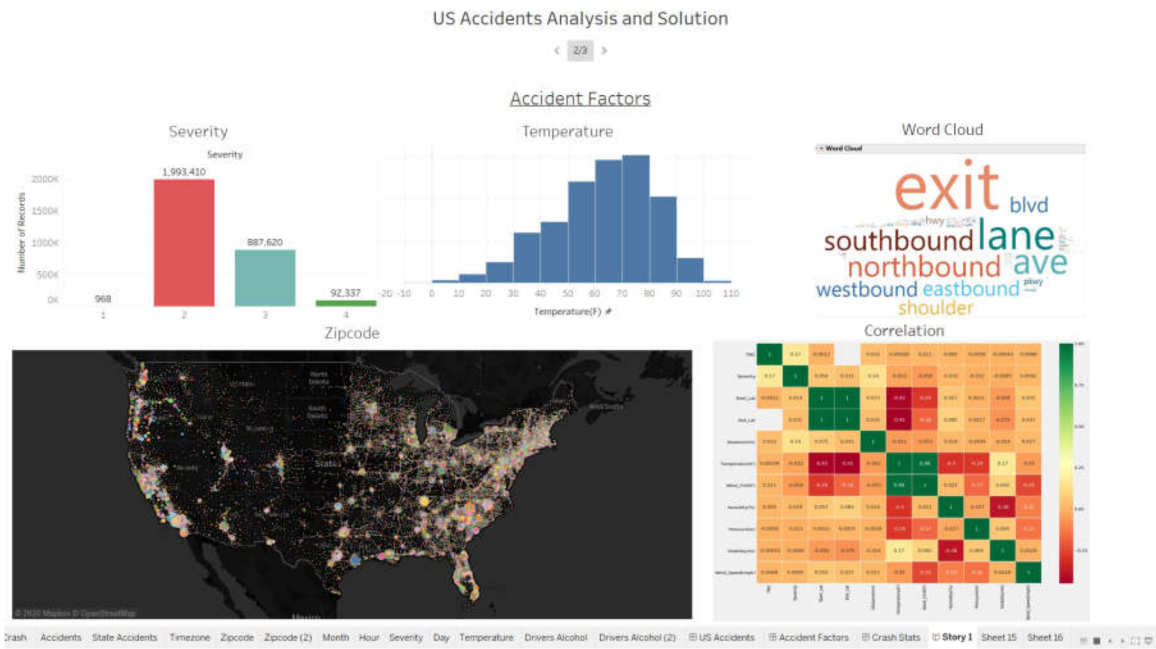

Figure 5-28:Accidents Factors

- Accidents Stats by State

  This dashboard focuses on death by each state, vehicles involved in the crash, and drivers with alcohol at the time of the accident. Also, it shows the trend for drivers with alcohol and cars involved in a collision from 2004-2018

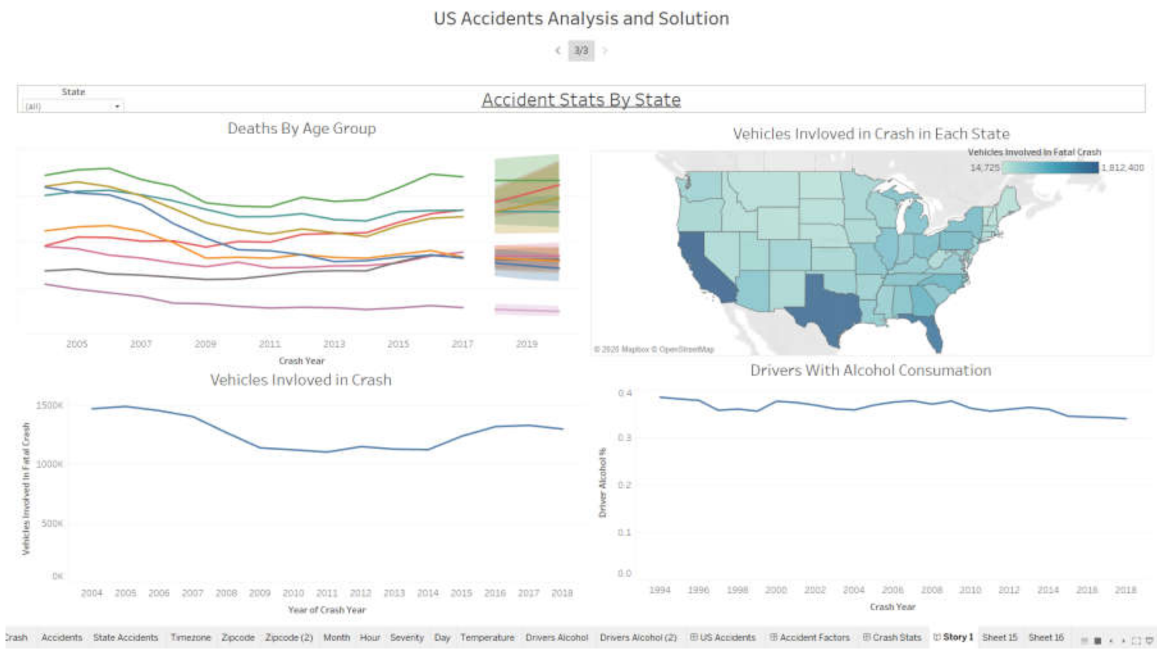

Figure 5-29:Accidents Stats by State

- The Solution to an Accident Based on Insights

  This dashboard will provide a recommendation based on insights found in the analysis.

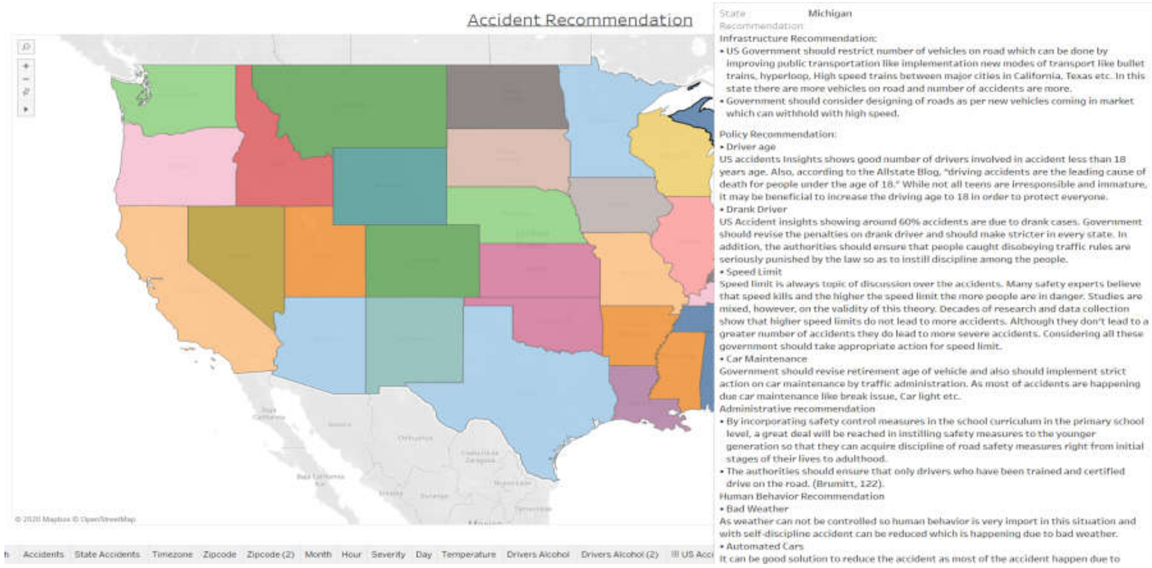

Figure 5-30:Accident Recommendation

CHAPTER FIVE

RECOMMENDATION

"Road accidents are unavoidable," says Livneh (Pablo, 2019). "But the risks of road accidents can be reduced by simple precautions and thinking ahead (Pablo, 2019)."

Based on the US accident analysis, there are many reasons for accidents like:

Table 6-0-1:Reason of the Accident Along with Recommendations

No	Reason for Accidents	Recommendation			
		Infrastructure	Policy	Administrative	Human Behavior
1	The increase in the number of drivers who disobey rules		✓	✓	
2	Drivers who drive their cars carelessly and disobedience of the rules		✓	✓	✓
3	Nature of the roads	✓		✓	
4	Owners not maintaining the vehicles properly			✓	✓
5	Bad weather				✓
6	Designing of the streets, for example, poor placements of the traffic controls	✓		✓	
7	Drunk Drivers		✓	✓	✓
8	High-Speed Vehicles	✓	✓	✓	

The solution to the above reasons falls under Infrastructure, Policy, Administrative, and human behavior (Self Discipline).

Infrastructure Recommendation

- The US Government should restrict the number of vehicles on roads, which can be done by improving public transportation through the implementation of new modes of transport like bullet trains, hyperloop, high-speed trains between major cities in California, Texas, etc. In these states, there are more vehicles on roads, and the number of accidents is higher.

Figure 6-1:Bullet Train (Nikada,2018)

Figure 6-2:Hyperloop (Quicler,2018)

- The government should consider the designing of roads as per new vehicles coming into the market, which can withhold high speed.

Policy Recommendation

Driver Age

US accident insights show a good number of drivers involved in accidents are less than 18 years of age. Also, according to the Allstate Blog, "driving accidents are the leading cause of death for people under the age of 18." While not all teens are irresponsible and immature, it may be beneficial to increase the driving age to 18 in order to protect everyone (Hammer,2019).

Drunk Drivers

US Accident insights show that around 60% of accidents are due to drunk cases. The government should revise the penalties on a drunk driver and should make them stricter in every state. Besides, the authorities should ensure that people caught disobeying traffic rules are severely punished by the law to instills discipline among the people.

Speed Limit

The speed limit is always a topic of discussion regarding the accidents. Many safety experts believe that speed kills, and the higher the speed limit, the more people are in danger. Studies are mixed, however, on the validity of this theory. Decades of research and data collection show that higher speed limits do not lead to more accidents. Although they do not lead to a more significant number of accidents, they do lead to more severe accidents. Considering this, the government should take appropriate action for speed limits.

Car Maintenance

The government should revise the retirement age of vehicles and should also implement strict action on car maintenance by traffic administration. Most accidents are happening due to car maintenance, like brake issues, car lights, etc.

Administrative recommendation

- The school curriculum should incorporate safety control measures at the primary school level, a great deal of awareness will be reached in instilling safety measures to the younger generation so that they can acquire the discipline of road safety measures right from the childhood

- The government authorities should make sure that only drivers who have been trained and certified to drive should be on the road.

Human Behavior Recommendations

Bad Weather

As the weather cannot be controlled, human behavior is essential in these situations. With self-discipline, accidents can be reduced.

Automated Cars

They can be an excellent solution to reduce accidents as accidents happen due to manual errors and repeated tasks. However, it will be essential to look at traditional and automated cars on the road to see what the impact will be. When every car is automatic, then it will reduce the number of accidents.

CHAPTER SIX

FUTURE WORK

In terms of my future work, I will be making this platform available to state governments and the public so it can be easily accessible. Different versions of this platform can be created, which can address specific problems of state government and people. This application needs to be launched on a website, and also live data will need to be fed to get a more up to date analysis.

Creating a machine learning model can help the public predict an accident location based on the source and destination location along with the date and time of travel. This type of prediction model can help reduce the number of accidents happening in the US. The prediction model can incorporate several neural network-based components that use a variety of data attributes, such as traffic events, weather data, points-of-interest, and time information.

CHAPTER SEVEN

CONCLUSION

Traffic accidents are a main public safety issue, with much research devoted to the analysis and prediction of these rare events. The study helped us to derive factors that are responsible for accidents. From this dataset, 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 travel. Also, it can help us to predict what are the accident-prone areas in each state. The analysis shows that the highest death is happening between the 20-35 age group, which is impacting the US economy. Most of the accidents occurring due to drunk driving. Finally, this study recommends infrastructure, Policy, Administrative, and Human Behavior changes, which can help to reduce US accidents.

APPENDIX A :

CODE

CODE

Word Cloud

In [26]:

```
more_stopwords=["accident", "due", "blocked", "Right", "hand"]
for more in more_stopwords:
    STOPWORDS.add(more)
# Generate a word cloud image
# lower max_font_size
wordcloud = WordCloud(stopwords=STOPWORDS, max_font_size=40, background_color="white").generate
(text)
plt.figure(figsize=(18, 10))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis("off")
plt.show
# Save the image in the img folder:
wordcloud.to_file("us_accidents_description.png")
```

Correlation

In [11]:

```
fig=sns.heatmap(df[['TMC','Severity','Start_Lat','End_Lat','Distance(mi)','Temperature(F)','Win
d_Chill(F)','Humidity(%)','Pressure(in)','Visibility(mi)','Wind_Speed(mph)']].corr(),annot=Tru
e,cmap='RdYlGn',linewidths=0.2,annot_kws={'size':15})
fig=plt.gcf()
fig.set_size_inches(18,15)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.show()
```

REFERENCES

Najjar, A., Kaneko, S., & Miyanaga, Y. (2017). Combining Satellite Imagery and Open Data to Map Road Safety. Thirty-First AAAI Conference on Artificial Intelligence.

Anderson D, Anderson J.(2004) Electric and hybrid cars: A history. McFarland, Jefferson;.

Bing Maps Dev Center - Bing Maps Dev Center. (2019). Retrieved from https://www.bingmapsportal.com/

Create live Word Clouds. (2019). Retrieved from https://www.mentimeter.com/features/word-cloud

Eisenberg, D. (2003, July 31). The mixed effects of precipitation on traffic crashes. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S000145750300085X?via=ihub

FARS Encyclopedia: States - Crashes and All Victims. (n.d.). Www-Fars.Nhtsa.Dot.Gov. Retrieved February 21, 2020, from https://www-fars.nhtsa.dot.gov/States/StatesCrashesAndAllVictims.aspx

Global status report on road safety (2009).Retrieved from https://www.who.int/violence_injury_prevention/road_safety_status/2009/en/

Rahman M (2015). Retrieved from https://rokibrahman.files.wordpress.com/2015/05/tableau-tutorial-for-beginners.pdf

Tamerius, J. D. (2016). Precipitation Effects on Motor Vehicle Crashes Vary by Space, Time, and Environmental Conditions. Retrieved from https://journals.ametsoc.org/doi/full/10.1175/WCAS-D-16-0009.1

Goniewicz, K., Goniewicz, M., Pawlowski W., Fiedor P. (2015). Road accidents in the early days of the automotive industry. Retrieved from https://www.academia.edu/19963502/Road_accidents_in_the_early_days_of_the_automotive_industry

Hammer, A. (2019, March 27). Legal Driving Age Around the World Driving ages around the world; which is best? Retrieved from https://blog.tirebuyer.com/legal-driving-age-around-the-world/

Chang, L.-Y. (2005, October 17). Analysis of freeway accident frequencies: Negative binomial regression versus artificial neural network. Retrieved from https://www.sciencedirect.com/science/article/pii/S0925753505000676

Official MapQuest - Maps, Driving Directions, Live Traffic. (2019). Retrieved from https://www.mapquest.com/

Motor-Vehicle-Crashes-Vehicle-Information-Three-Ye/xe9x-a24f. (2019). Accessed: 2020-02-10.

New York Traffic Accidents. (2019). Retrieved from https://data.ny.gov/Transportation/

Nikada. (n.d.). View of Modern High-speed Bullet trains in Beijing south train... Retrieved from https://www.gettyimages.com/detail/photo/modern-high-speed-trains-in-beijing-china-royalty-free-image/1030988672?adppopup=true

Pablo Q. Causes and Solutions of Car Accidents (2019) - 1117 Words: Report Example. Retrieved from https://ivypanda.com/essays/causes-and-solutions-of-car-accidents/

Python Introduction. (n.d.). Retrieved from https://www.w3schools.com/python/python_intro.asp

Chen, Q., Song, X., Yamada, H., & Shibasaki, R. (2016). Learning Deep Representation from Big and Heterogeneous Data for Traffic Accident Inference. Conference on Artificial Intelligence.

Quicler, C. (n.d.). A full-scale passenger Hyperloop capsule is presented by Hyperloop... Retrieved from https://www.gettyimages.com/detail/news-photo/full-scale-passenger-hyperloop-capsule-is-presented-by-news-photo/1044539698?adppopup=true

Kumar S., Toshniwal D. (1970, January 1). A data mining framework to analyze road accident data. Retrieved from https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0035-y

Seattle (WA) Traffic Accidents. (2019). Retrieved from http://data-seattlecitygis.opendata.arcgis.com/datasets/5b5c745e0f1f48e7a53acec63a0022ab_0.

Moosavi, S., Samavatian, M. H., Nandi, A.,Parthasarathy, S. & Ramnath R. (2019, July 1). Short and Long-term Pattern Discovery Over Large-Scale Geo-Spatiotemporal Data. Retrieved from https://dl.acm.org/doi/10.1145/3292500.3330755

The Economic and Societal Impact of Motor Vehicle Crashes (2010). US Department of Transportation. [PDF file]. Retrieved from https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812013

Tuovila, A. (2020, January 29). Forecasting Definition. Retrieved from https://www.investopedia.com/terms/f/forecasting.asp

US-Accidents: A Countrywide Traffic Accident Dataset. (2019, December 31). Retrieved from https://smoosavi.org/datasets/us_accidents?cv=1

Wolfgang K, Rybicki J, Mauve M (2007). On the nature of inter-vehicle communication. Communication in Distributed Systems ITG-GI Conference. VDE; 2007

Xu W., Hu S, Cui Y, Ma Y (2005). Study on Evaluation of Harmonious Development Between Transportation and Economy. JBJU (Social Sciences Edition) ; 2:004

Yuan Z., Zhou X., Yang T., Tamerius J, and Mantilla R.(2017). Predicting traffic accidents through heterogeneous urban data: A case study. In Proceedings of the 6th International Workshop on Urban Computing , Halifax, NS, Canada, Vol. 14.