DEPARTMENT OF MATHEMATICS COMPUTER SCIENCE & ENGINEERING TECHNOLOGY

FEATURES EXPLORATIONS OF US ROAD ACCIDENTS FROM 2016 TO 2021

by

Bay Bryan

A Thesis submitted to the Graduate Faculty of
Elizabeth City State University
in partial fulfillment of the
requirements for the Degree of
Master of Science in
Applied Mathematics.

December

2022

APPROVED BY

Julian D. Allagan, Ph.D. Committee Chair	Kenneth L. Jones, Ph.D. Committee Member
Mohamed Elbakary, Ph.D. Committee Member	Gabriela H. Del Villar, Ph.D. Committee member

ABSTRACT

This thesis explores factors that influence traffic accidents focusing on location, severity, time, road condition, and weather factors. Using a nationwide traffic accident dataset found in Kaggle which contains Application Programming Interface (API) data from 2016 to 2021, we visualize these important features of the data. Further, our work focuses on providing answers to more than a dozen data exploratory questions which shine important lights on factors that may contribute to the rise of the number of accidents, along with accidents severity.

DEDICATION

I would like to dedicate this thesis to my parents Pam and Russell Bryan. My parents have always encouraged me to pursue my dreams and have helped in the process. They were there for my highs and lows of my studies and pushed me to not give up. I would also like to dedicate this thesis to my boyfriend and friends. My boyfriend Conner Wilson and friend Ellee Scott were willing and did indeed drop everything they were doing in our small town of Ohio to move down to North Carolina with me so that I could pursue my masters degree at Elizabeth City State University. They both helped significantly in the movement down and emotionally with all the nerves that can come when moving to a new state. Without their love and support it would have been difficult to make the decision to move and be able to pursue another degree. My boyfriend who stayed in North Carolina with me, for the entirety of the academic year, I am especially grateful for. He took on more household duties during the "busier" school months always willingly because of his unconditional love for me, which I cherish. All of my close friends I am thankful for who encouraged me through this and made it known how proud of me they are.

ACKNOWLEDGEMENT

I would first like to acknowledge two Professors I had in the process of receiving a bachelor's degree Professor Welsh at Zane State College and Dr. Elliot Paquette at The Ohio State University (now at McGill University), they both helped in any possible way they could when I had questions taking there class and encouraged me to pursue an even higher degree. Dr. Paquette even wrote a recommendation for me to Elizabeth City State University. While at Elizabeth City State University, I would also like to acknowledge Professor Shatoya Covert, going into my masters degree (and being the first of my family to do so) I was nervous and not sure what to expect. My landlord had connections to Professor Covert and told me to reach out to her and I am glad that I did. Professor Covert set my mind at ease and went over and above to make me feel welcome at ECSU, she even explained the process of beginning to write a thesis to me since this was my first one. I had the honor of working with Dr. Julian Allagan on my thesis. His knowledge helped me in many ways on this thesis and I have learned so much from him. Dr. Allagan is an incredible professor and mentor. He even went out of his way to find me several job opportunities when I arrived to North Carolina which I am very grateful for and the jobs I enjoyed tremendously. Dr. Allagan going above and beyond for his students inspires me and I strive to be an educator like him. I would also like to thank Dr. Kenneth L. Jones who is a great professor at thoroughly explaining different topics. Dr. Jones even personally reached out to me when I got accepted into Elizabeth City State University to answer any of my questions. I would like to thank Dr. Talukder as he was always ready to help other students and I on any questions we may have had and came to class always happy and excited to teach. Lastly, I would like to thank Dr. Dipendrda Sengupta who would thoroughly explain anything I was having trouble with understanding and would help in any way he could so his students could be sucessful in his course. Thank you again to all of my professors, I appreciate everything you all have taught me.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

1.1 Background and Overview

Developed in 2016, a dataset was established to disclose information about vehicle accidents in the United States. The data is to be used only for Research purpose, under the Creative Commons Attribution-Noncommercial-ShareAlike license. This research covers accident records from A Countrywide Traffic Accident Dataset collected from February 2016 to December 2021. The information used is posted in real-time, using multiple traffic API's. There are around 2.8 million accident records currently in the data-set [11].

Using Python notebook we explore and interpret the data via several visualization tools from Python libraries such as matplotlib, seaborn, plotly, geopandas, etc. The results include locations of accidents, and bar plots to illustrate accident frequencies and times. In Chapter 1, we introduce the literature background, explain the meaning of several key variables such as 'Severity', 'Time', 'Road Condition', etc.

In Chapter 2, we describe some of the variables in the data and their distributions by addressing several research questions as listed, next. In Chapter 3, we explore the different contributing factors and changes we found in year 2021.

Our research relies on the use of Python programming language (IDE Jupyter Notebook) to explore the data by answering the following questions:

Question n° 1. *Which city has the highest no. of road accidents in US (2016-2021)?*

Question n° 2. *In the past 6 years how many accidents happened in the city from no.1?*

Question n° 3. *How many from the top 10 cities with the most no. of accident cases is from the state, California?*

Question n° 4. *How many cities in the US have only 1 accident record in the past 6 years?*

Question n° 5. *Which state has the highest no. of road accidents in the past 6 years?*

Question n° 6. *Which state has the lowest no. of road accidents in the past 6 years?*

Question n° 7. *What time zone region of US has the highest no. of road accident cases in the past 6 years?*

Question n° 8. *In the last 6 years which highway has the highest road accident records?*

Question n° 9. *How many streets have only 1 accident record in the past 6 years?*

Question n° 10. *The impact on traffic was Moderate (Severity-2) in how many percent of cases?*

Question n° 11. *The majority of road accidents have impacted traffic flow for how many hours?*

Question n° 12. *Around what percent of road accidents occurred in December?*

Question n° 13. *What percent of road accidents occurred during the weekend?*

Question n° 14. *What is the most deadliest accident hour?*

Question n° 15. *How many road accidents happened near a crossing?*

Question n° 16. *What is the temperature/ humidity/ air pressure/ wind chill/ wind speed/ visibility range during the most amount of accidents?*

Question n° 17. *What type of weather had the highest amount of accident cases?*

Answers for each question is provided and a figure with an interpretation is given, where appropriate. We close our thesis with Chapter 5 with possible future research that will extend some preliminary results on features important using Neural Network and other machine learning techniques.

1.2 Literature Review

Extensive research on vehicle accidents gave valuable insight on why accidents are happening and where they are most likely to occur. By using data analysis techniques

to portray the statistics in an easy-to-interpret format the research information can be used to create significant changes in the future to help lower the number of vehicle accidents that occur each year in the United States. By reviewing literature and articles on vehicle accidents the aim is to examine the information from a statistical perspective and provide insight on how to reduce accidents. From our reviews, we found five influential factors:

1. location
2. severity
3. time
4. road condition
5. weather

We provide more details regarding each factor, next.

1.2.1 Location

Crashes occur for a number of reasons. In 2021, Dolphin Technologies analyzed 3.22 million car trips between 2018 and 2019 and found that 25 percent of all accidents happened during the first three minutes of driving. We will explore what cities and locations which are more prone to accidents. The law office of Deborah M. Truscello reported, for the year 2021, the top reasons why accidents are occurring in overpopulated areas. They listed distracted driving, parked vehicles on narrow roads, and vehicles parked on the side of all different types of roads. Other reasons listed are impaired driving, speeding, and inexperienced drivers [13]. We will also research, according to Movement Mortgage, how a recent shift in population movement is affecting where accidents are happening and show data that backs up this claim [3].

1.2.2 Severity

As innovation progresses toward an increase in car traffic concerns, other factors associated with vehicle accidents also sparks interest, which includes severity. Although

not much research has recently been done on accident severity, it is an important topic we will explore as it give insight on traffic and expectations of delays in the future.

1.2.3 Time

The 2021 Global Traffic Scorecard recorded that Americans lost 3.4 billion hours due to congestion in 2021 which is down by 42% from pre-COVID times. This means that the average American driver lost 36 hours due to congestion. The effect of COVID-19's impact on traffic remained constant in the year 2021 with reduced traffic times, the volume of people on the road being cut down, and less downtown traffic. Cities such as New York and Chicago lost the most amount of time in traffic, but the amount of time lost was still down over 20% compared to pre-COVID times [6].

1.2.4 Road Condition

Factors such as wind speed, precipitation, fog, pavement temperature, pavement condition, and water level can contribute to causing incidents from road conditions. Wind speed causes trouble with visibility and lane obstruction from blown snow, dust, and debris. Different types and rates of precipitation can cause visibility issues, pavement friction, and lane obstruction. Fog impacts visibility distance. Pavement temperature and condition can cause infrastructure damage and the condition also leads to pavement friction. Lastly, according to the Department of Transportation's list of road weather variables water level impacts the roadway by lane submersion [14].

1.2.5 Weather

Weather plays a major variable in vehicle accidents causing roadway, traffic flow, and operational impacts. According to data from the National Highway Traffic Safety Administration, an average of around 21% of accidents each year are weather related. Weather related crashes are defined by those that occur during adverse weather conditions (rain, sleet, snow, fog, etc.) or on slick pavement. On freeways, light rain or

snow can reduce speed up to 13% while heavy rain decreases the average speed by up to 16%. In heavy snow, the average freeway speeds decline by as much as 40%. Overall, it has been estimated that 23% of delays on highways across the United States are due to snow, ice, and fog [14].

Chapter 2 Data Exploration

2.1 Data Description and Attributes

A country-wide traffic data-set is used which covers the United States. This is a large scale publicly available database of accident information called US-Accidents. The data (1.15GB, `.../input/us-accidents/US_Accidents_Dec21_updated.csv`) is being continuously collected dating back to February 2016. The data uses several providers including APIs (Application Programming Interface) that broadcast traffic events captured by entities such as the US and State Department of transportation, law enforcement agencies, traffic cameras and traffic sensors within the road-networks. Currently, there are over 2.8 million accident records in the data-set. Each accident record consists 47 attributes including Severity, Location, Time, Weather, etc. See Table 2.1, for more details.

#	Attribute	Description	Type
1	ID	This is a unique identifier of the accident record.	S
2	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).	N
3	Start_Time	Shows start time of the accident in local time zone.	N
4	End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	N
5	Start_Lat	Shows latitude in GPS coordinate of the start point.	N

6	Start_Lng	Shows longitude in GPS coordinate of the start point.	N
7	End_Lat	Shows latitude in GPS coordinate of the end point.	N
8	End_Lng	Shows longitude in GPS coordinate of the end point.	N
9	Distance(mi)	The length of the road extent affected by the accident.	N
10	Description	Shows natural language description of the accident.	S
11	Number	Shows the street number in address field.	N
12	Street	Shows the street name in address field.	S
13	Side	Shows the relative side of the street (Right/Left) in address field.	S
14	City	Shows the city in address field.	S
15	County	Shows the county in address field.	S
16	State	Shows the state in address field.	S
17	Zipcode	Shows the zipcode in address field.	N
18	Country	Shows the country in address field.	S
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	S
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	S
21	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	N
22	Temperature(F)	Shows the temperature (in Fahrenheit).	N
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	N
24	Humidity(%)	Shows the humidity (in percentage).	N
25	Pressure(in)	Shows the air pressure (in inches).	N
26	Visibility(mi)	Shows visibility (in miles).	N
27	Wind_Direction	Shows wind direction.	S
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	N
29	Precipitation(in)	Shows precipitation amount in inches, if there is any.	N

30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	S
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	S
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	S
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	S
34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	S
35	Junction	A POI annotation which indicates presence of junction in a nearby location.	S
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	S
37	Railway	A POI annotation which indicates presence of railway in a nearby location.	S
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	S
39	Station	A POI annotation which indicates presence of station in a nearby location.	S
40	Stop	A POI annotation which indicates presence of stop in a nearby location.	S
41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	S
42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby loction.	S
43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	S

44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	S
45	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	S
46	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	S
47	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	S

Table 2.1: Table of Code Names Called and What They Mean

Type Key:

S- String

N- Numeric

2.1.1 Location

To develop the graphs, panda was called in and a data-frame was created of cities and their corresponding accident cases. The library mpatches was also brought in to customize the plots.

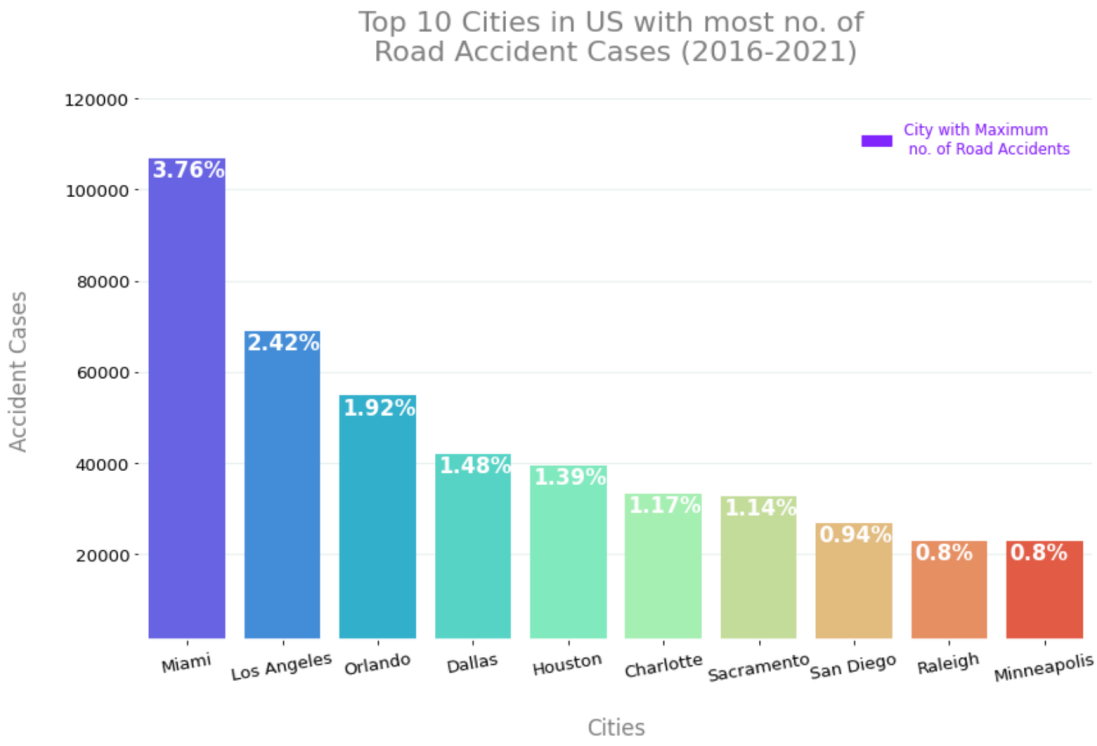


Figure 2.1: Location of Incidents (2016-2021)

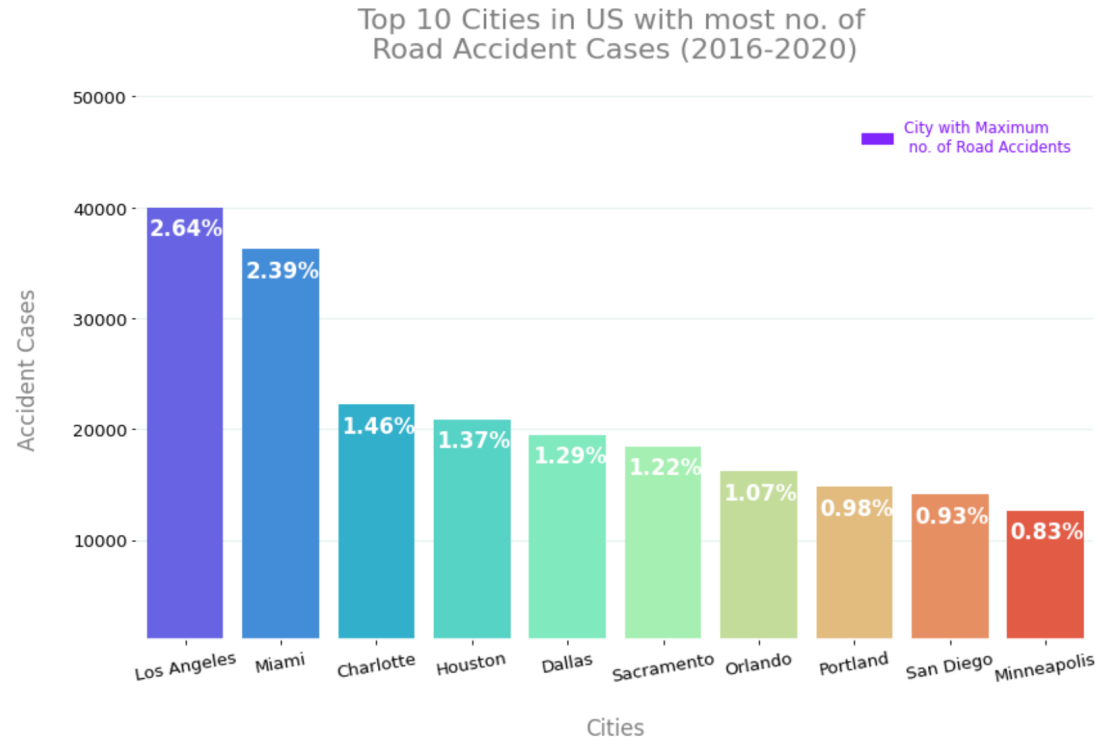


Figure 2.2: Location of Incidents (2016-2020)

The accidents are recorded over 11,679 cities as per the data-set up to 2021. Miami has the highest number of road accidents in the US (3.76%). Within the past 6 years

on average 17,828 road accidents happen in Miami each year, in 24 hours around 49 accidents occur. Los Angeles is the second highest with (2.42%). Around 16% accident records from the past 6 (2016-2021) years account for only 10 of the major cities listed out of 10,657 cities in the data-set. Three out of the top 10 cities with the highest number of accidents are located in California.
Compared to data from 2016-2020: Los Angeles had the highest recorded accidents with 2.64% and Miami falling slightly behind with 2.39%.

2.1.1.1 City Cases Percentage

Below is a table of the data used and information found on accident cases over 11,679 cities in the US.

Data Code	**Number of Cities**	**Percent Overall**
city_cases_percentage(1, '=')	1110 Cities	9.5%
city_cases_percentage(100, '<')	8727 Cities	74.71%
city_cases_percentage(1000, '<')	11185 Cities	95.75%
city_cases_percentage(1000, '>')	494 Cities	4.23%
city_cases_percentage(5000, '>')	71 Cities	0.61%
city_cases_percentage(10000, '>')	34 Cities	0.29%

Table 2.2: Data Code used and Resulting Outcome about Accidents within Cities

Data Interpretation:

1. 9.5% (1,110 Cities) in the US have only 1 accident record in the past 6 years.
2. Almost 75% (8,727 Cities) have less than 100 total number of road accidents.
3. 95.75% (11,185 Cities) have less than 1,000 total number of road accidents.
4. 4.23% (494 Cities) have more than 1,000 total number of road accidents.
5. 71 Cities (0.61%) have had over 5,000 road accidents in the past 6 years.
6. Only 34 Cities (0.29%) have had over 10,000 road accidents in the past 6 years.

2.1.1.2 State Analysis

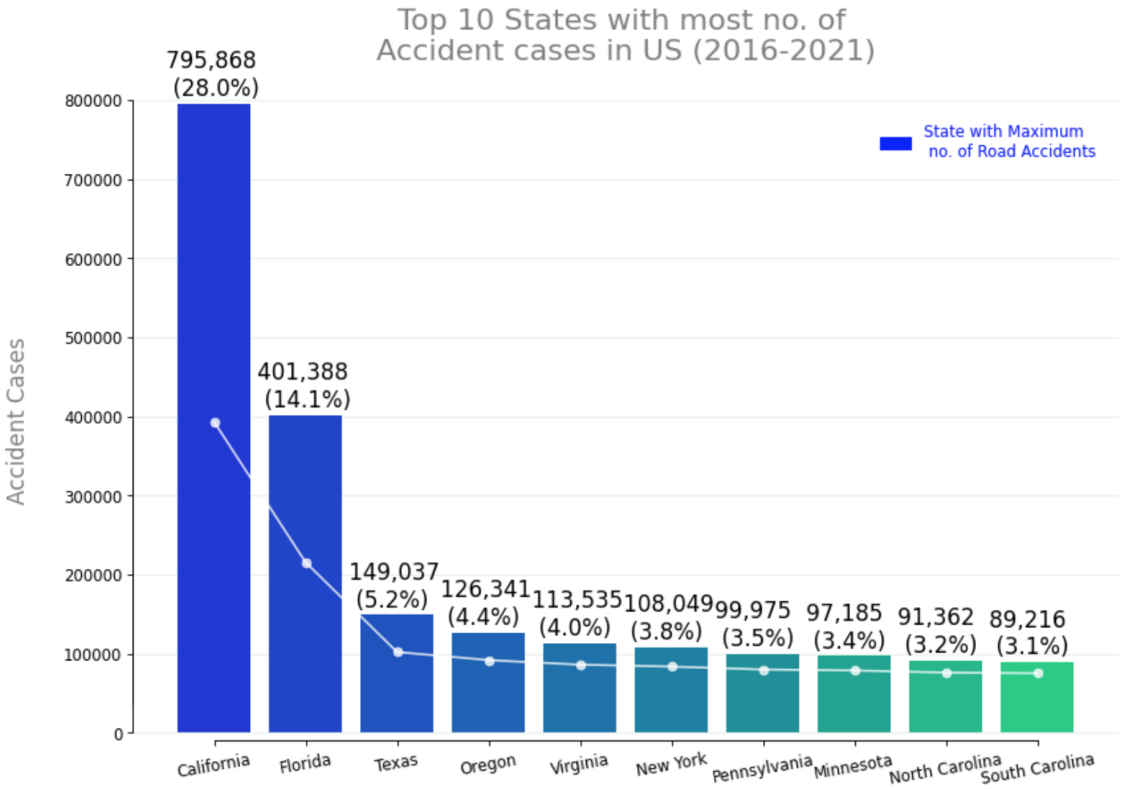


Figure 2.3: Incidents by State

To create the plots, in which we could easily review the top ten states, geopandas was called in to read the file while a geolocator was used to search OpenStreetMap data by location. A dictionary was created of all the US states and territories, excluding Alaska, and a data-frame of states and their corresponding accident cases were created. Lastly, a function "convert" was used to switch the state code with their actual corresponding name.

We can see from the graph, Figure 2.3 that California has the highest accident cases out of all the states. It has almost double as many as Florida with Florida being the second highest (14.1%). About 28% of all accident records from the past 6 years are from California, this implies that about 15 accidents happen in California per hour. Compared to data from 2016-2020: California still remained the state with the highest number of recorded accidents with 448,833 recorded accidents up to this point meaning in the year 2021 alone California had 347,035 accidents occur.

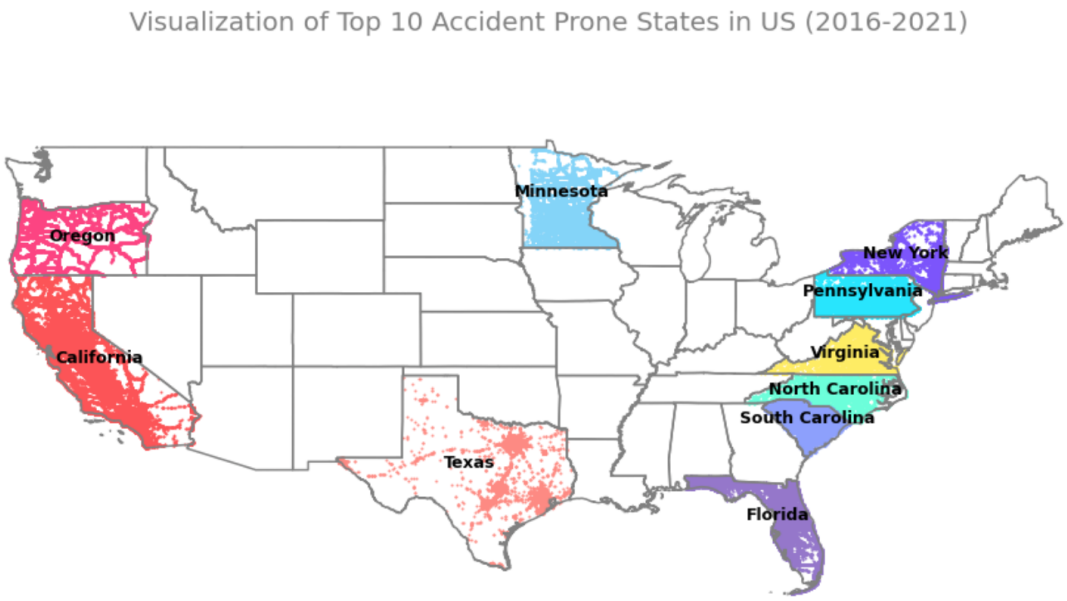


Figure 2.4: Points of Major State Incidents

This map (Figure 2.4) depicts the incidents happening in the top 10 states and pinpoints the major location of road accidents happening in these states. To create the map, the library geopandas needed to be called in along with geometry to check for a point within a polygon and to be precise of the points depending on longitude and latitude of the map.

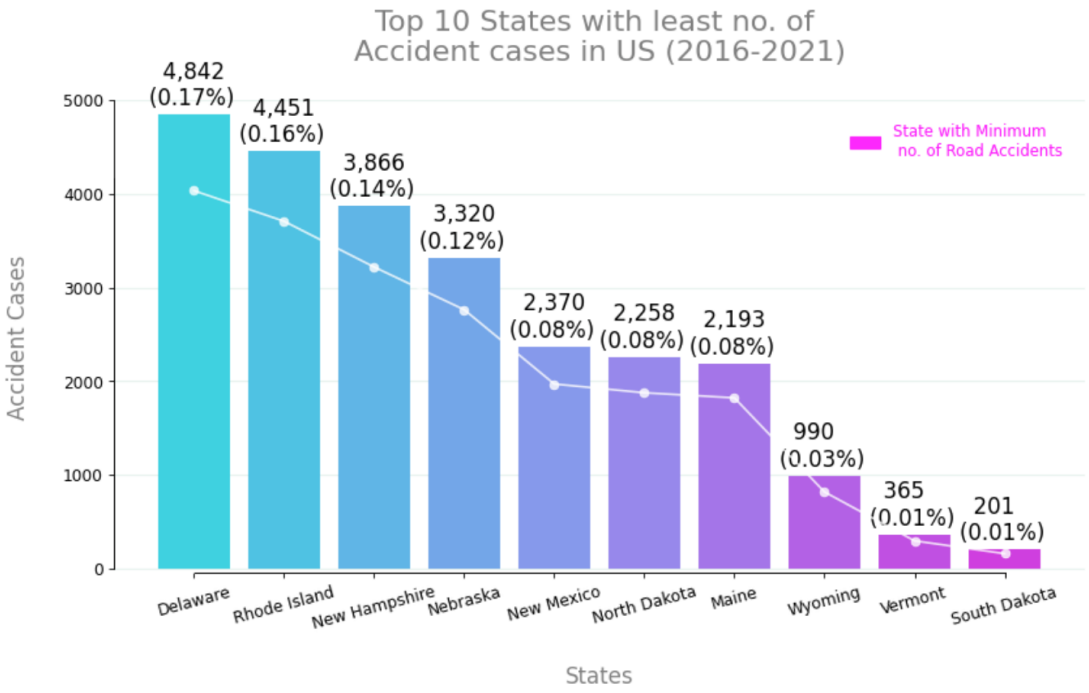


Figure 2.5: States with the Least Amount of Incidents

Since we have covered the states with the most number of road accidents we can now evaluate the top ten states with the least amount of road accidents Figure 2.5 above shows these top ten states. As we can see South Dakota has the least amount on record with 201 accidents on record for the past 6 years.

2.1.1.3 Timezone Analysis

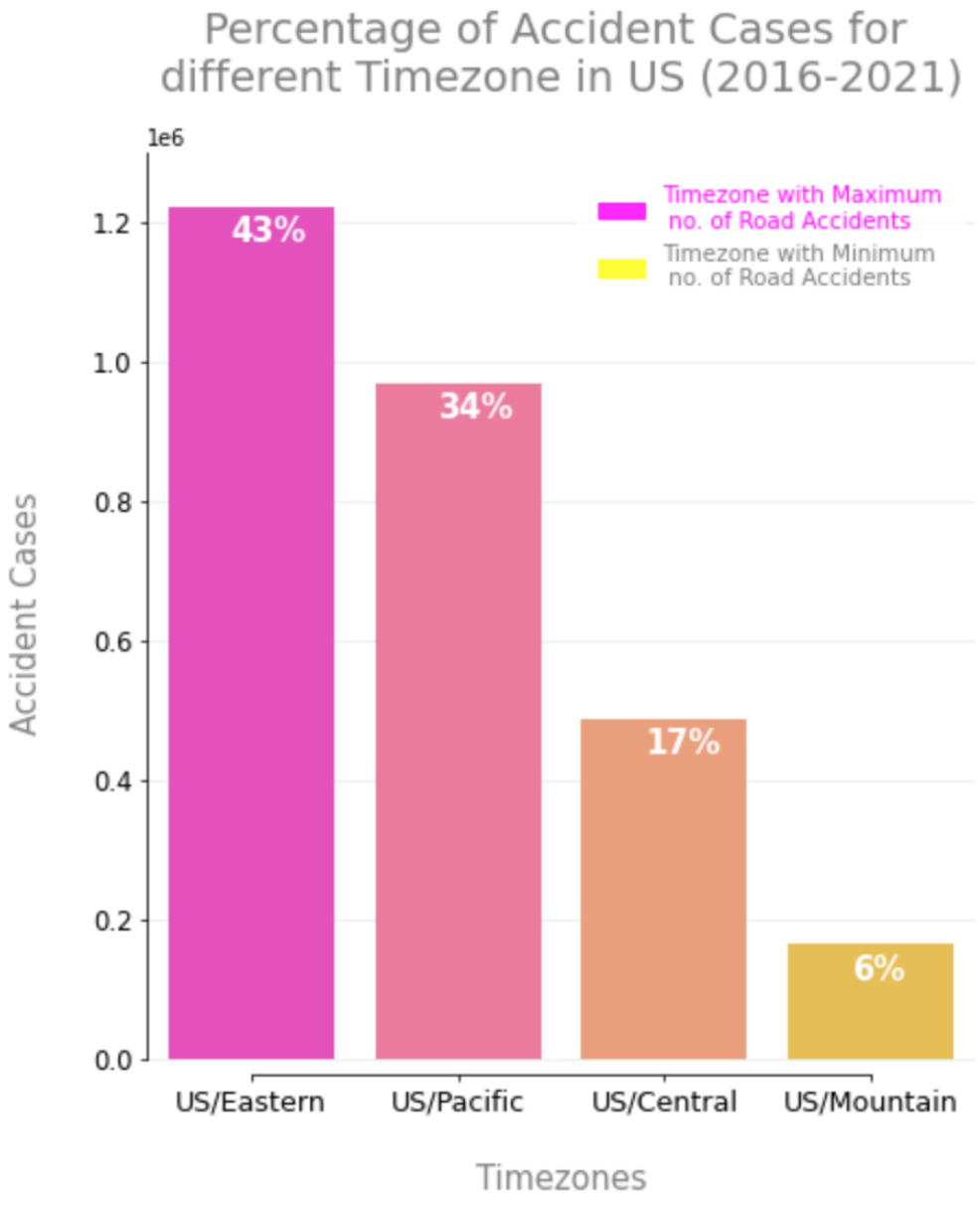


Figure 2.6: Accident Cases for Different Timezones in the US

Eastern timezone has the most amount of accidents with 43% while while Mountain timezone has the least amount (with 6%) in the last 6 years.

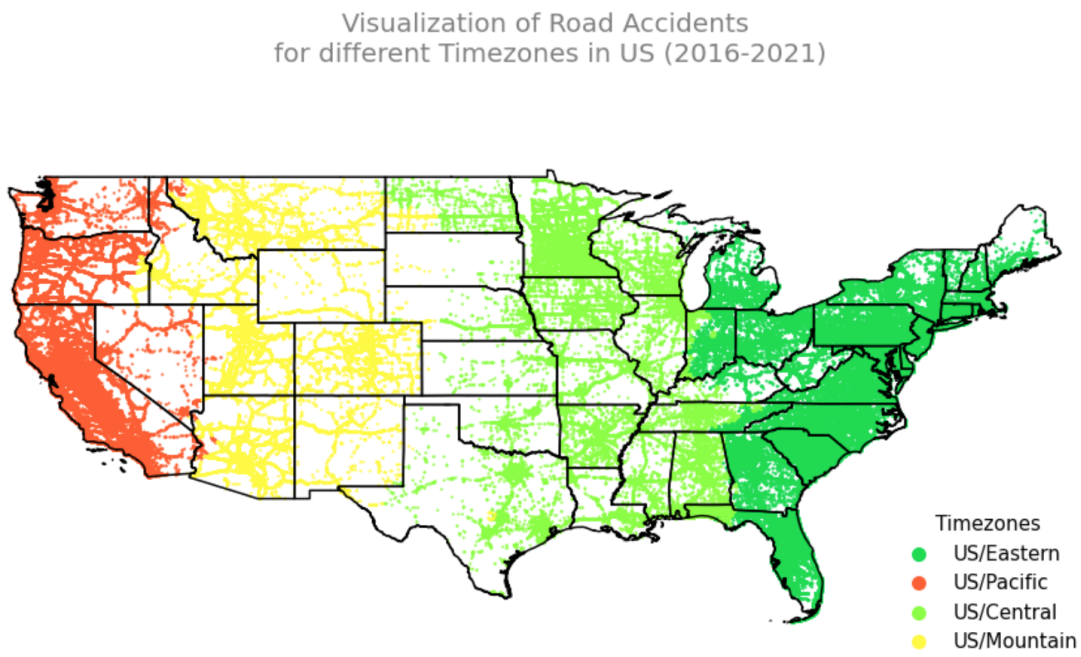


Figure 2.7: Visualization of Road Accidents within Timezone

2.1.1.4 Street Analysis

We were able to pull data about major highway accidents and what street these accidents occurred. Using top_ten_streets_df we can create a plot with the ten streets that have the highest amount of accidents.

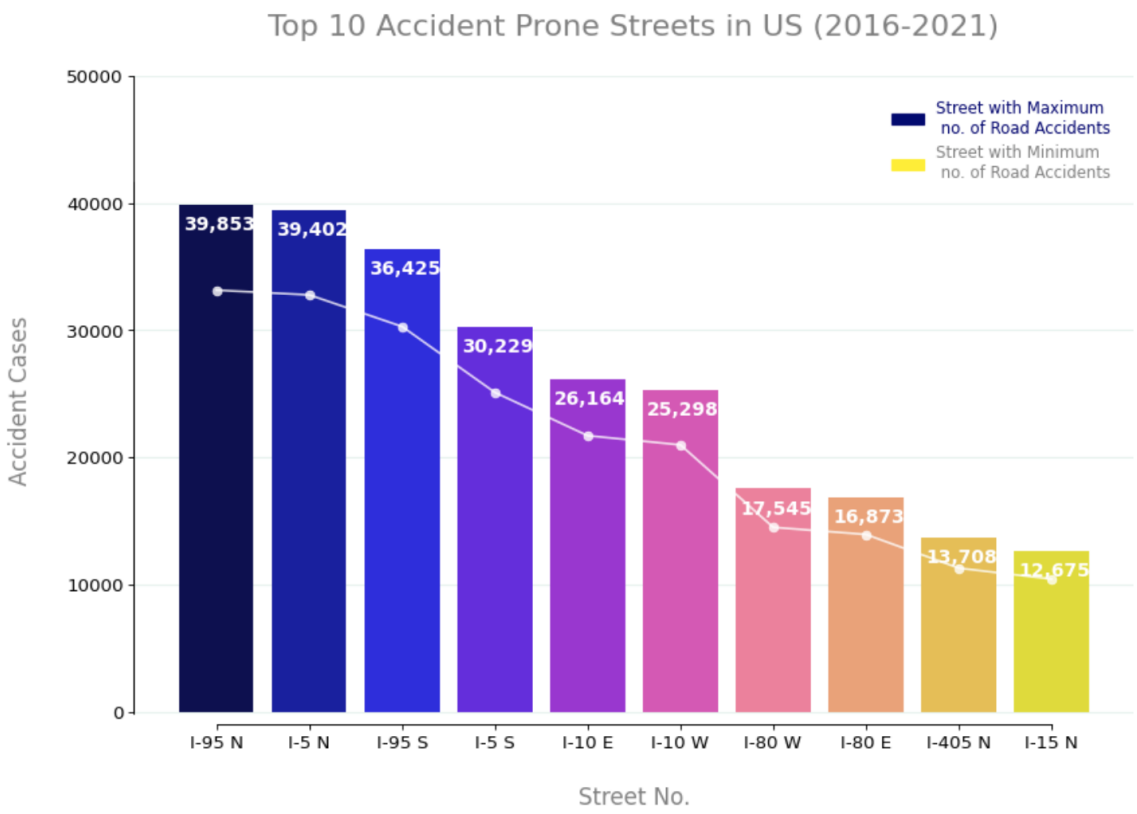


Figure 2.8: Top 10 Streets with the Highest Amount of Accidents

Looking at Figure 2.8 we can see in the last 6 years (2016-2021), I-95 N has had the most amount of accidents with an average of 18 accidents per day.

Compared to data from 2016-2020: I-5 N had the highest accidents recorded with an average of 14 accidents per day. (I-95 N came in second with an average of 12 accidents per day.)

Next, we will use the data on streets to see the amount of road accidents happening per street from 159,650 streets enlisted below:

Data Code	**Number of Streets**	**Percent Overall**
street_cases_percentage(1, '=')	64,154 Streets	40.18%
street_cases_percentage(100, '<')	156,364 Streets	97.94%
street_cases_percentage(1000, '<')	159, 325 Streets	99.8%
street_cases_percentage(1000, '>')	325 Streets	0.2%
street_cases_percentage(5000, '>')	56 Streets	0.04%

Table 2.3 continued from previous page

Table 2.3: Data Code used and Resulting Outcome about Accidents on Different Streets

This information shows there are 64,154 Streets from the data with only 1 accident that has occurred. 97.94% have had less than 100 accidents occur on them while 99.8% (159,325) of streets have had 1,000 accidents or less. Only 325 streets in the US recorded in the past 6 years have had over 1,000 accidents occur on that street specifically. 56 streets total have had over 5,000 accidents occur on them. Since we searched the amount of streets with over and under 1,000 accidents those combined give us our total amount of streets within the data-set which is 159,650 streets total.

2.1.2 Severity

Below is a chart that shows the severity of an accident, the severity meter is given a scale between 1-4. The number 1 indicates the least impact on traffic going all the way up to number 4 which has the most impact on traffic (i.e., a long delay). Severity from different sources may differ slightly on their overall impact on traffic. To show a severity representation, a data-frame was created of severity and corresponding accident cases. The figure was separated into four sections based on severity levels.

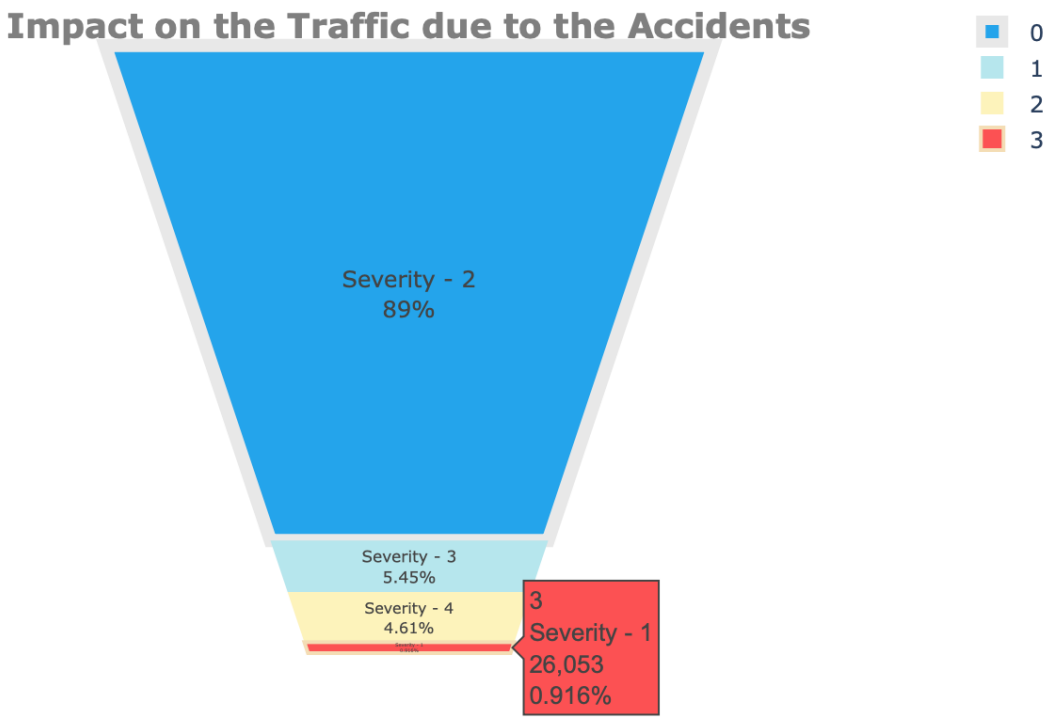


Figure 2.9: Severity Levels

As we can see from 2.9 the highest amount of cases with 89% (2,532,991 Accidents) is severity level 2 which is a slight delay in traffic (moderate) from the accident. The lowest amount of cases with 26,053 equivalent to 0.916% is severity level 1 which is little to no delay in traffic. 4.61% of cases is severity level 4 which has a highly severe level of impact on traffic.

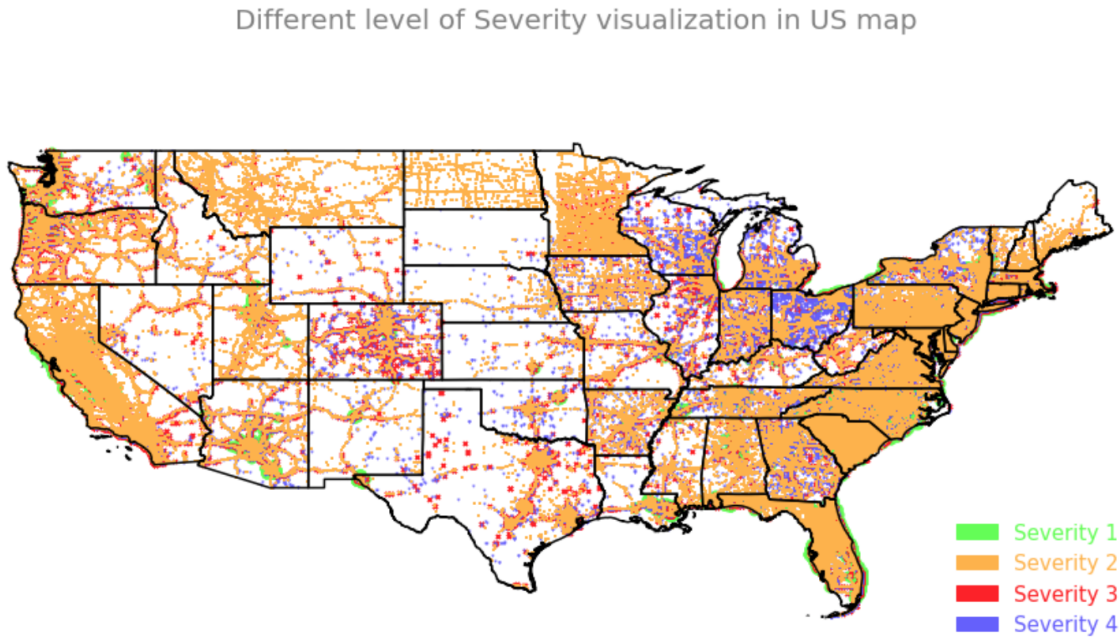


Figure 2.10: Visualization of Severity Levels

2.1.3 Time

For the different times we use Start_Time and End_Time to indicate the starting and ending time of an incident within their local time zones.

2.1.3.1 Accident Duration

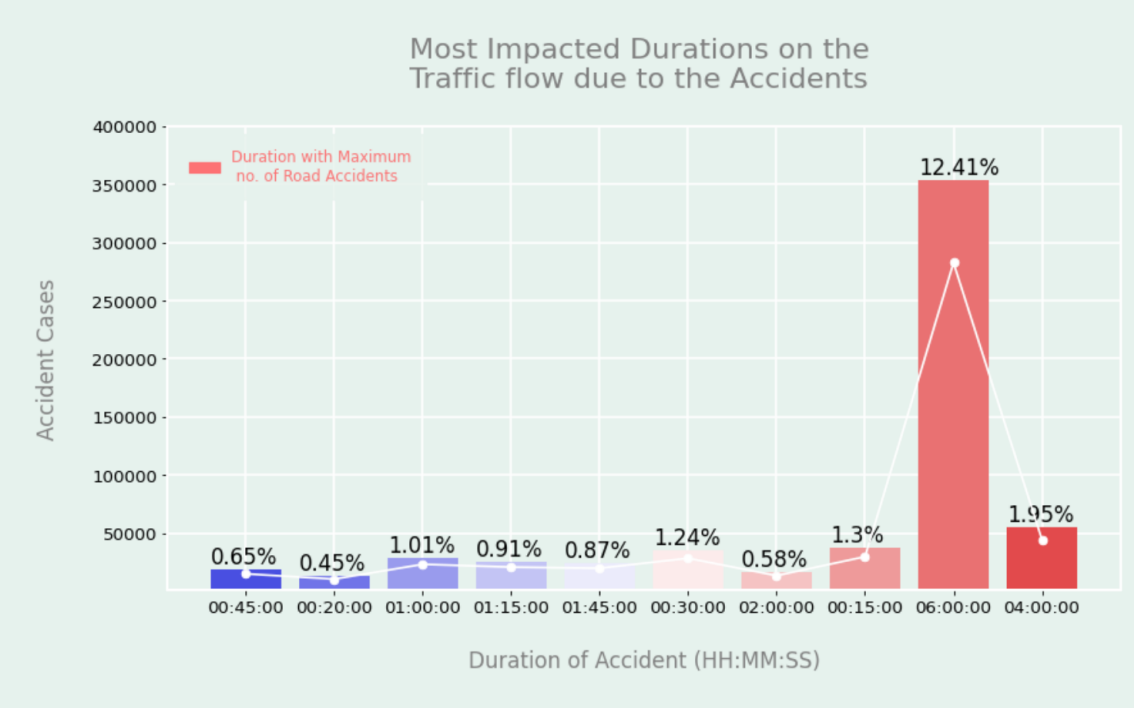


Figure 2.11: Accident Duration Analysis

Figure 2.11 shows how long an accident has an impact on traffic flow for. As shown on the graph, the majority, 12.41% has an impact on traffic for a duration of 6 hours. The least amount of accidents 0.45% has an impact on traffic flow for 20 minutes.

2.1.3.2 Year Analysis

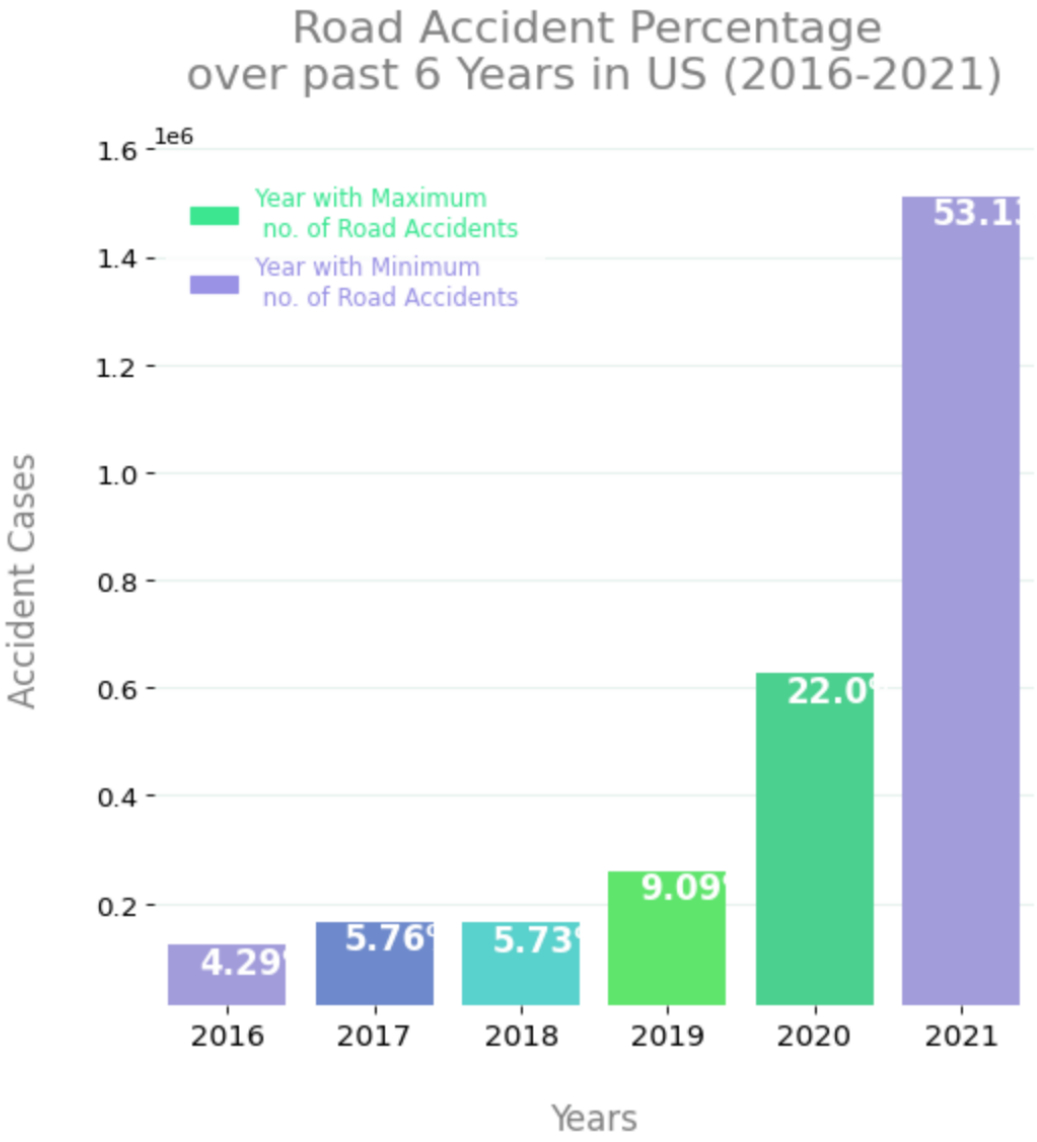


Figure 2.12: Amount of Accidents over the Years

To create Figure 2.12, mpatches and a ticker was used to easily customize the graph and to add marks for every 200,000 road accidents that occurred.

As shown in Figure 2.12 the amount of accidents in 2021 has more than doubled since the previous year 2020. 75% of accidents happening from 2016-2021 happened only within the past 2 years (2020,2021). The year 2016 has the least amount of road accidents with 122,024 accidents happening that year compared to 2021 where 1,511,745 accidents occurred. Below is a visualization of the road accidents over the past 6 years.

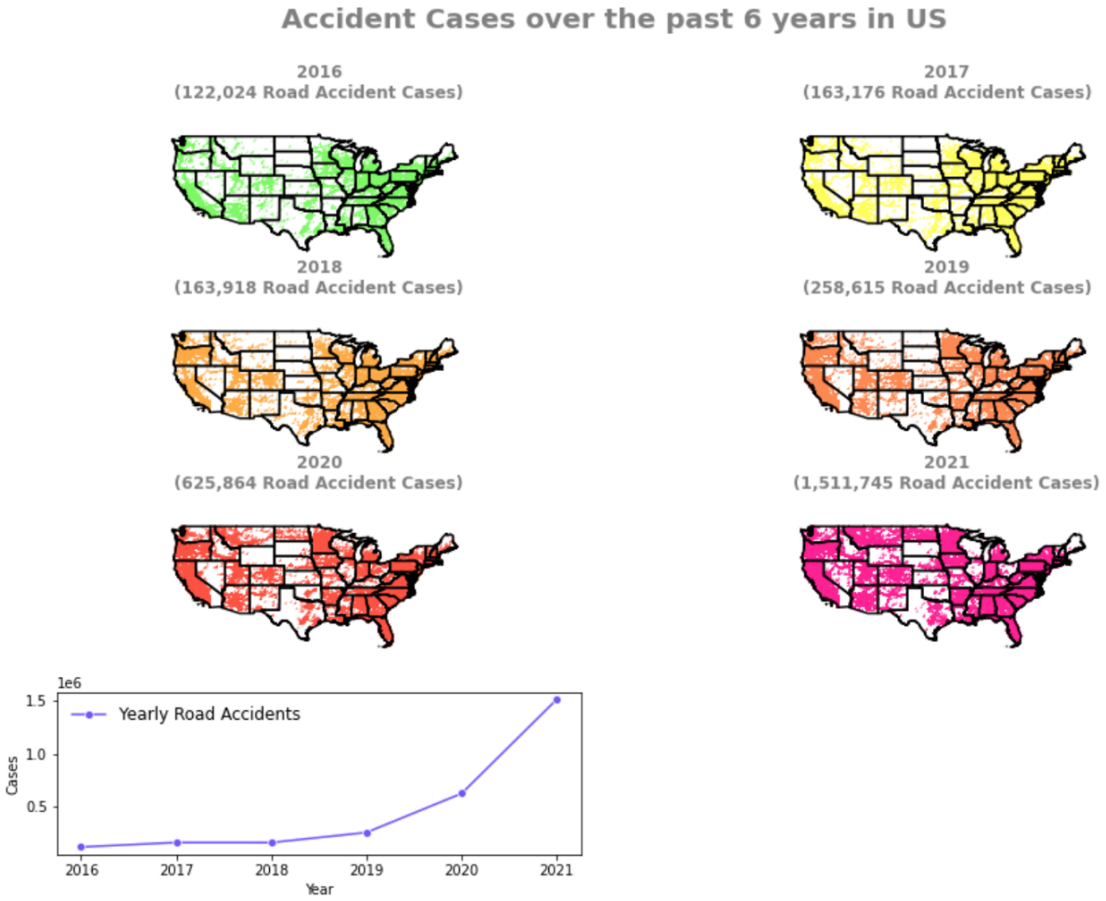


Figure 2.13: Records of Accidents from (2016-2021)

2.1.3.3 Month Analysis

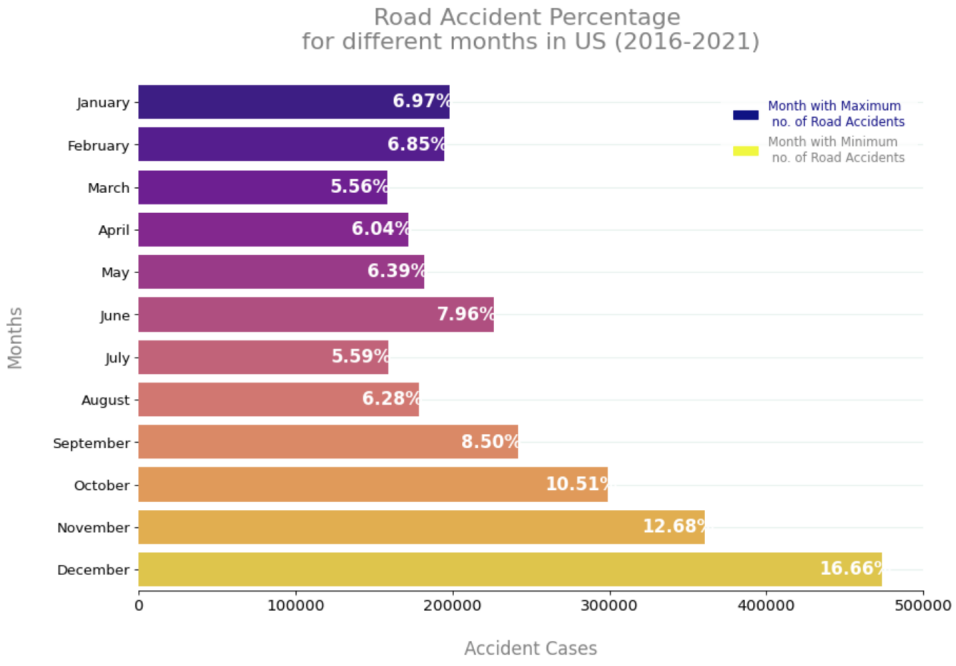


Figure 2.14: Accidents Each Month from (2016-2021)

Figure 2.15 shows the amount of accidents that have happened on each different month from 2016-2021. December has had the most amount of accidents with 16.66%, while March has the least amount of accidents with 5.56% of accidents happening overall. Notice how about 40% of accidents that took place occurred within a 3 month span, October to December which are the main months in the US that represent the transition period from Autumn to Winter.

2.1.3.4 Day Analysis

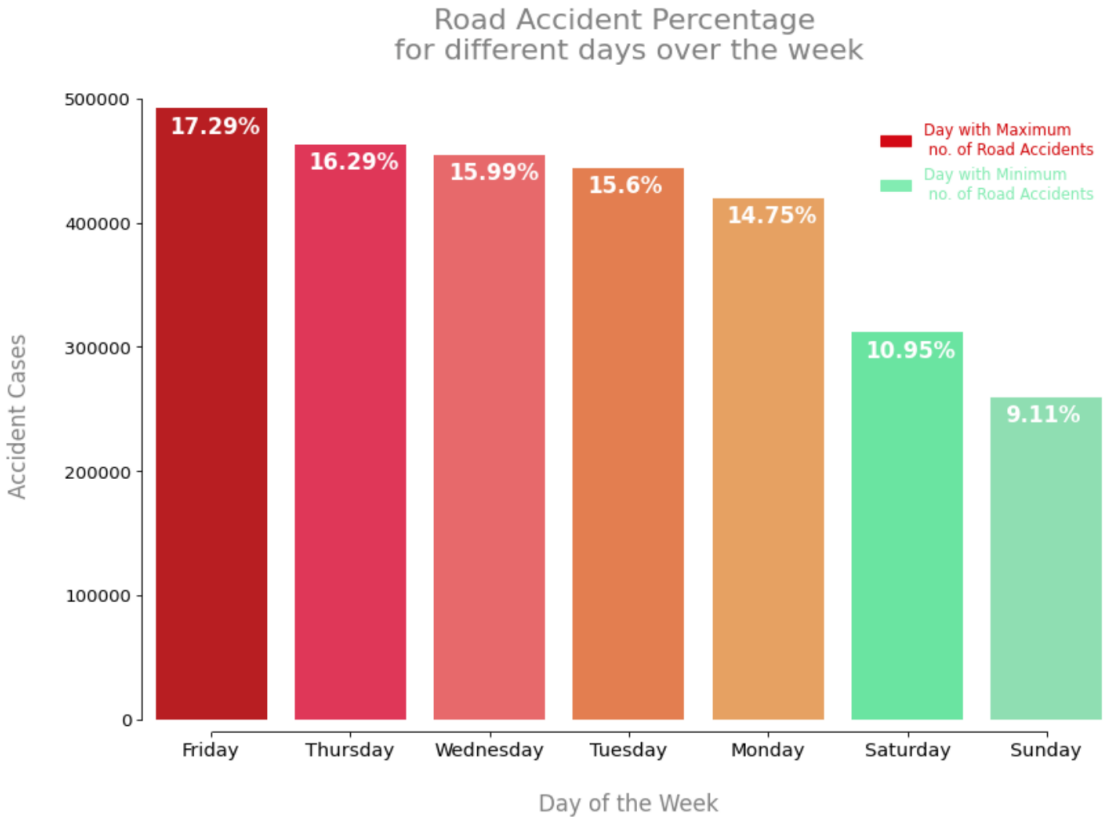


Figure 2.15: Accidents Each Day from (2016-2021)

Figure 2.15 shows what day each accident has occurred from 2016-2021. The greatest amount of accidents occur on a Friday with 17.29% of accidents, while the least amount of accidents occur on a Sunday (9.11%). Notice how working days have almost 2 times higher percentage than weekends. Only 20% of road accidents happen during the weekend.

Compared to data from 2016-2020: Thursday had the greatest amount of road ac-

cidents with 17.02% and Wednesday falling slightly behind with 16.87% and Friday being the third highest.

2.1.3.5 Hour Analysis

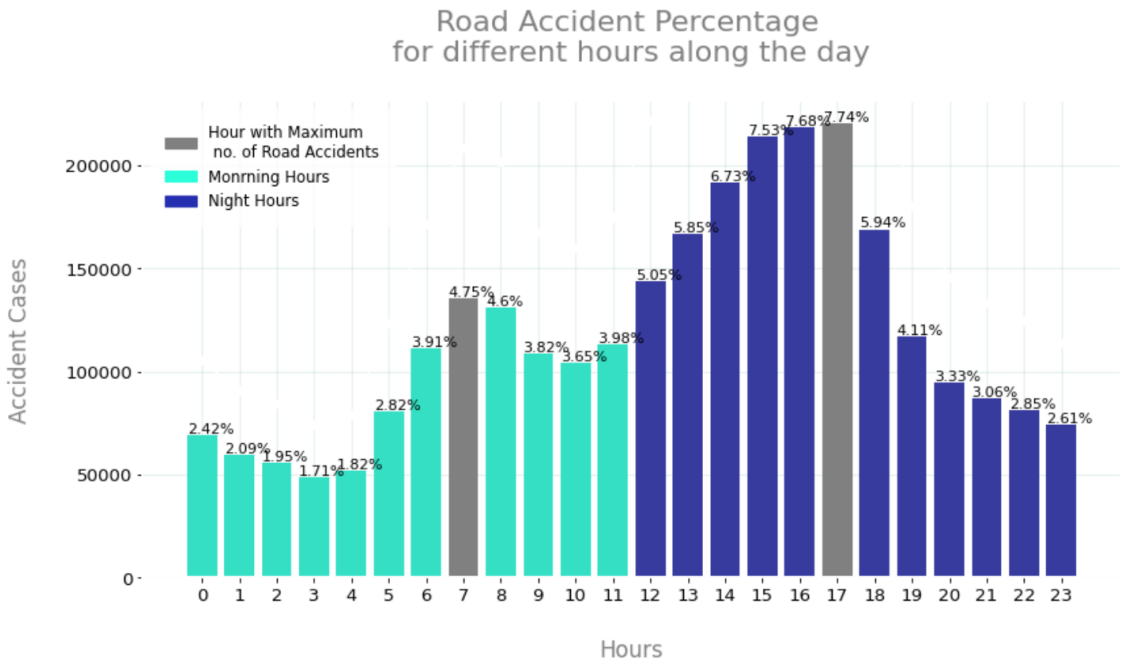


Figure 2.16: Amount of Accidents that Occur Each Hour

From the Figure above (2.16) around 17% of road accidents occur between 6:00AM and 9:00AM. 21% of road accidents occur between 3:00PM and 6:00PM. Notice how these hours average around heading to work and leaving from work. The most deadliest hour in the morning is 7:00AM (with 4.75%) which is a main "office-going" hour while the most deadliest time in the night is 5:00PM (with 7.74%) which is also a main "office-leaving" hour.

Compared to data from 2016-2020: 8:00AM was the most deadliest morning hour while the most deadliest night hour remained the same at 5:00PM. This could hint to a possible meaning that in the year 2021 the majority of people began work at an average of one hour earlier.

2.1.4 Road Condition

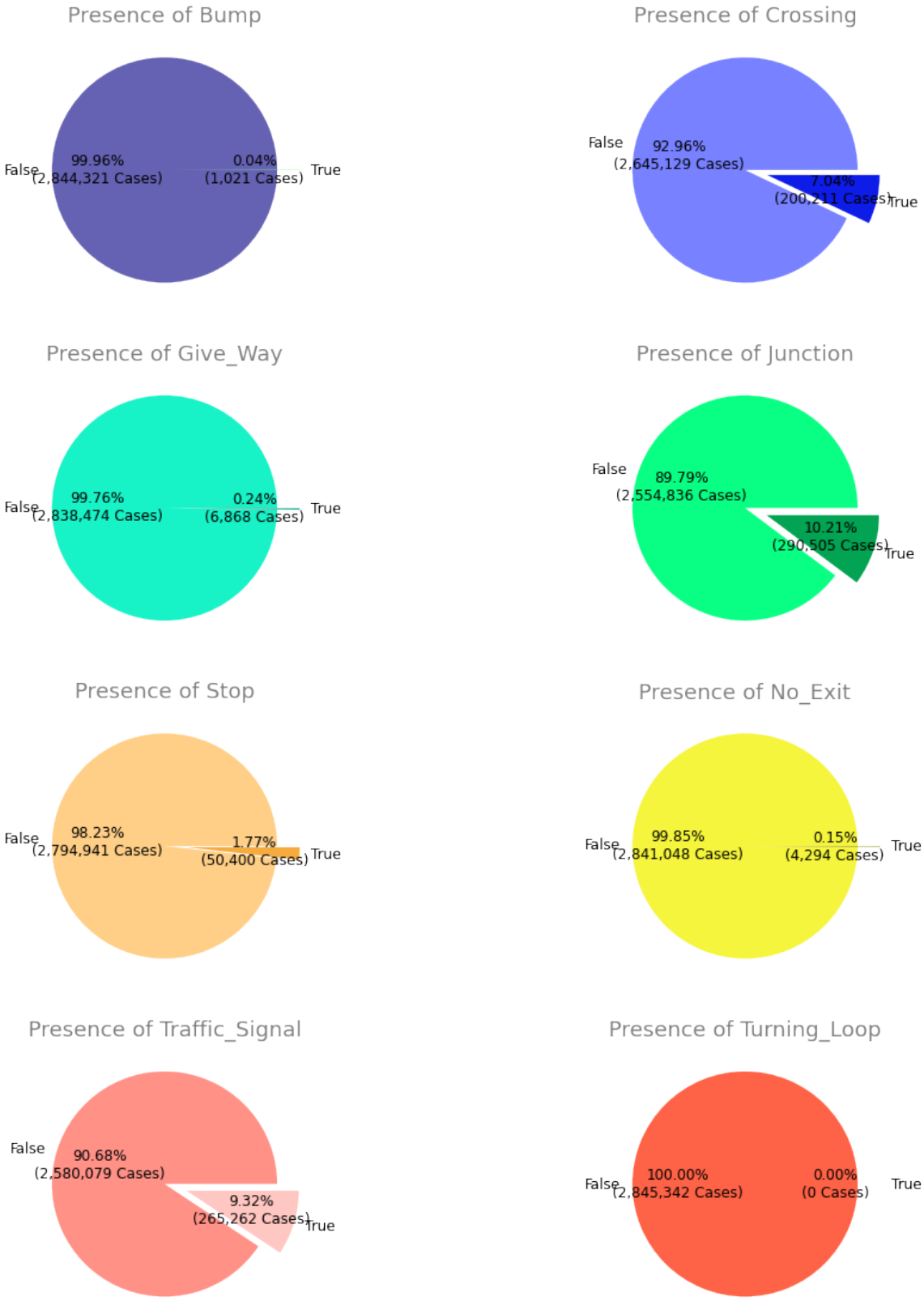


Figure 2.17: Presence of Different Conditions when an Accident Occurred

From the information above, (Figure 2.17) in almost every case (99.96%) the presence of a bumper was absent. In 7.04% of cases a crossing was at the incident site. 10.21% of cases had the presence of a junction. There were almost no stop signs at the presence of an accident only 1.77%. Only 0.15% of cases had the presence of a No Exit sign. A traffic signal was present in 9.32% of cases. A turning loop was not present in any cases of the given data.

2.1.5 Weather

Before discussing the newest data on weather condition accidents, it is useful to see how these different conditions could have an effect on causing an incident. According to the Department of Transportation, specific variables in weather can have these type of impacts:

Road Weather Variables	Roadway Impacts	Traffic Flow Impacts	Operational Impacts
Air temperature and humidity	N/A	N/A	Road treatment strategy (e.g., snow and ice control) Construction planning (e.g., paving and striping)
Wind speed	Visibility distance (due to blowing snow, dust) Lane obstruction (due to wind-blown snow, debris)	Traffic speed Travel time delay Accident risk	Vehicle performance (e.g., stability) Access control (e.g., restrict vehicle type, close road) Evacuation decision support
Precipitation **(type, rate, start/end times)**	Visibility distance Pavement friction Lane obstruction	Roadway capacity Traffic speed Travel time delay Accident risk	Vehicle performance (e.g., traction) Driver capabilities/behavior Road treatment strategy Traffic signal timing Speed limit control Evacuation decision support Institutional coordination
Fog	Visibility distance	Traffic speed Speed variance Travel time delay Accident risk	Driver capabilities/behavior Road treatment strategy Access control Speed limit control
Pavement temperature	Infrastructure damage	N/A	Road treatment strategy
Pavement condition	Pavement friction Infrastructure damage	Roadway capacity Traffic speed Travel time delay Accident risk	Vehicle performance Driver capabilities/behavior (e.g., route choice) Road treatment strategy Traffic signal timing Speed limit control
Water level	Lane submersion	Traffic speed Travel time delay Accident risk	Access control Evacuation decision support Institutional coordination

Table 2.4: How Different Variables in Weather Impact the Roadway, Traffic Flow, and Operation of a Vehicle

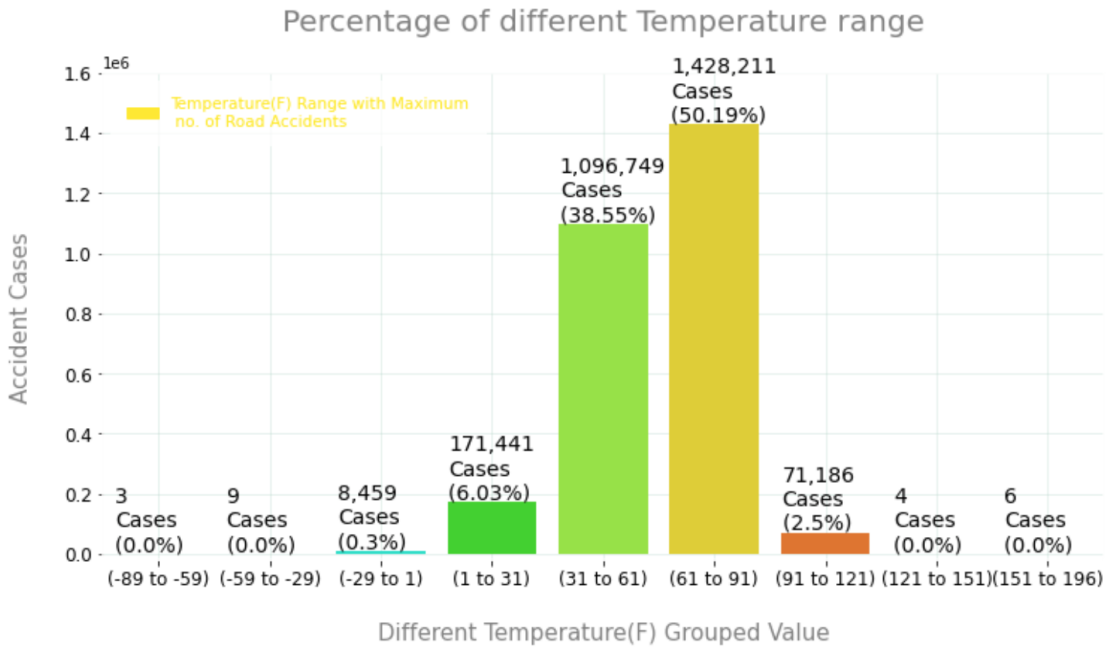


Figure 2.18: Accident Percentage During Different Temperatures (in Fahrenheit)

Above is a graph of when accidents occurred during different temperatures. There were 0 accidents that occurred in the United States (possibly excluding Alaska) below -29 degrees Fahrenheit and above 121 degrees Fahrenheit. Over half of the accidents that occurred were when the temperature range was from 61 to 91 degrees.

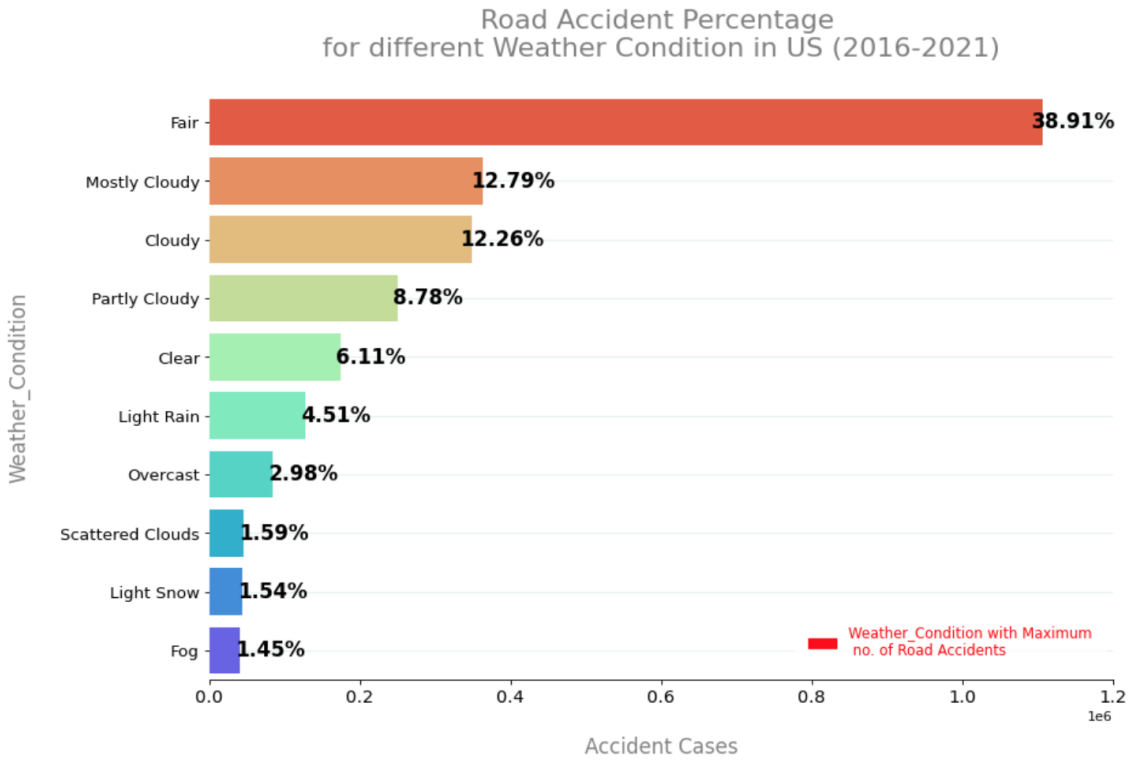


Figure 2.19: Weather Conditions in the US (2016-2021)

In most accident cases (38.91%) the weather condition was fair during the time of the accident. In almost 13% of cases the weather was mostly cloudy.

2.1.5.1 Humidity

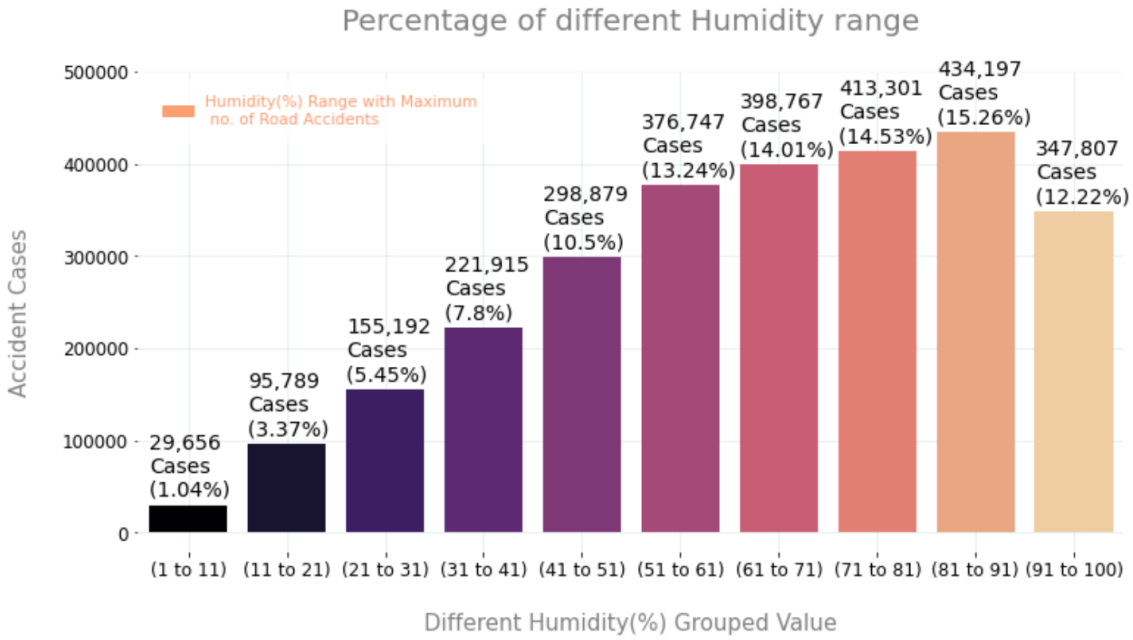


Figure 2.20: Accidents During Different Humidity's

In the maximum amount of cases the humidity range was between 81% to 91%. We can see a correlation, as the humidity rises the accident percentage rises as well until the humidity gets to the 91% mark.

2.1.5.2 Air Pressure

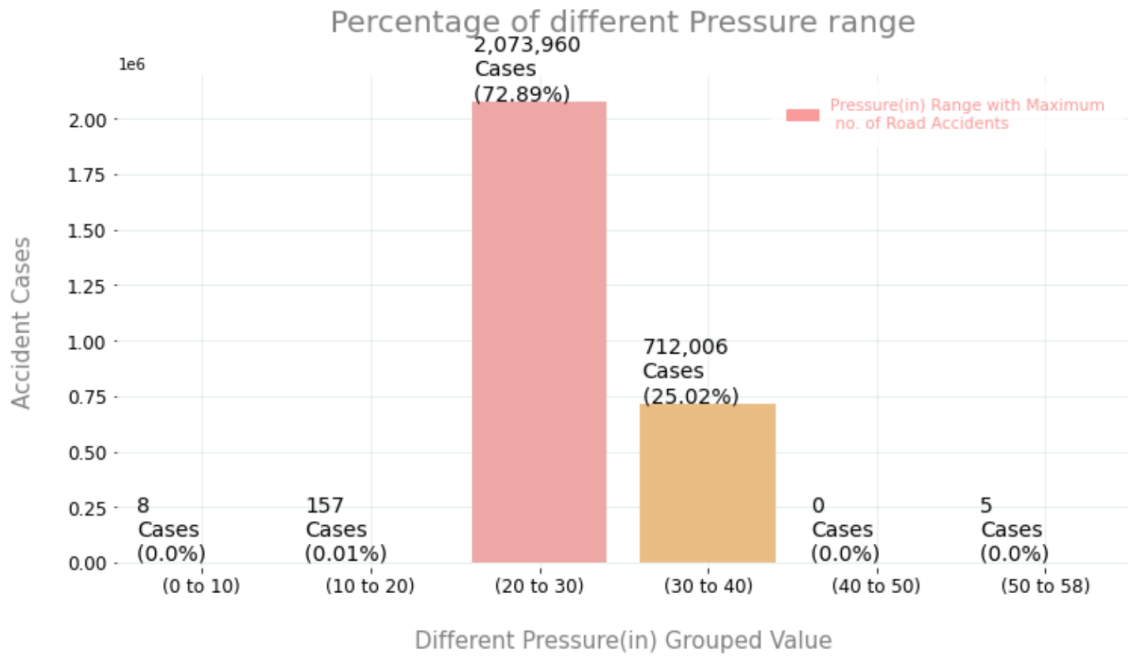


Figure 2.21: Accidents During Different Air Pressure's

In Figure 2.21 we are able to obtain the data of the different air pressures in the atmosphere at the time of the given accident. As we can see in 72.89% of cases of road accidents that air pressure is between 20 to 30 inches.

2.1.5.3 Wind

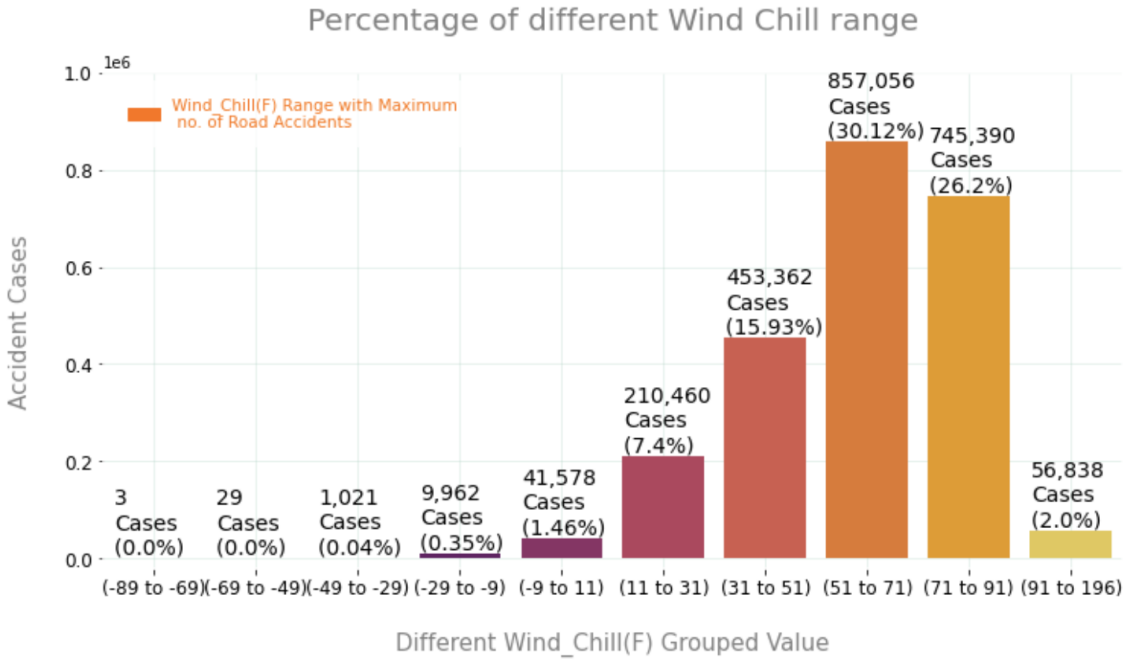


Figure 2.22: Accidents at the Time of Different Wind Chills

In the majority of cases, 30.12% of road accidents, the wind chill was between 51 to 71 degrees Fahrenheit.

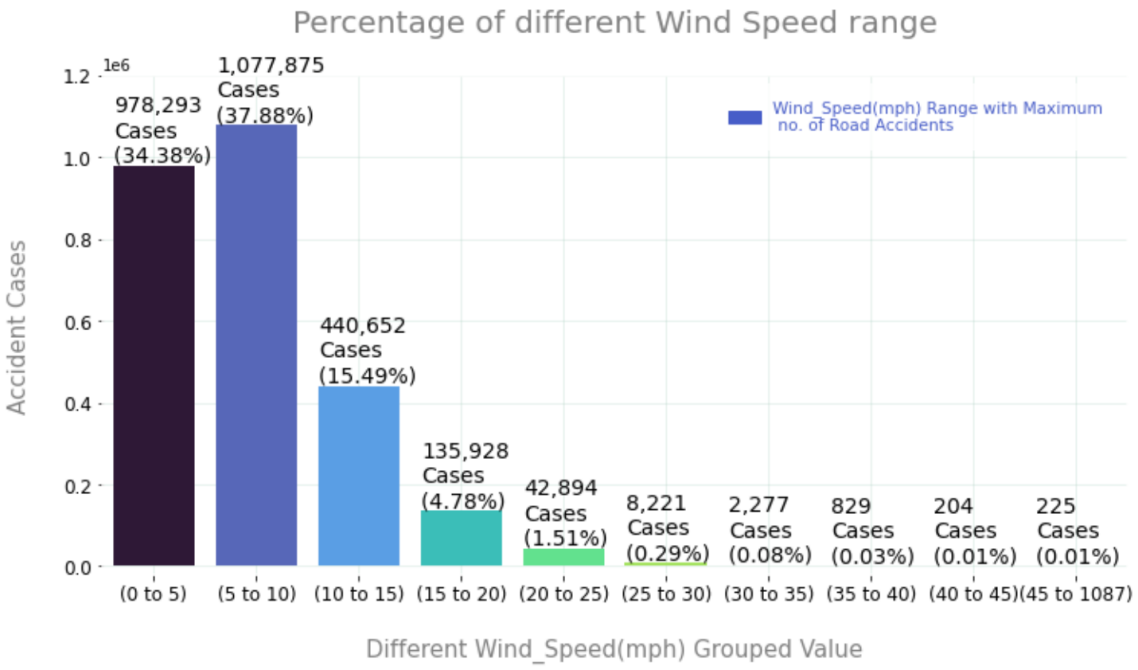


Figure 2.23: Accidents at the Time of Different Wind Speeds

As seen above, the majority of accidents happened (37.88%) when the wind speed

range was low between 5 to 10 miles per hour. From this data, we can assume that the majority of accidents from 2016 to 2021 were not affected by the speed of wind.

2.1.5.4 Visibility

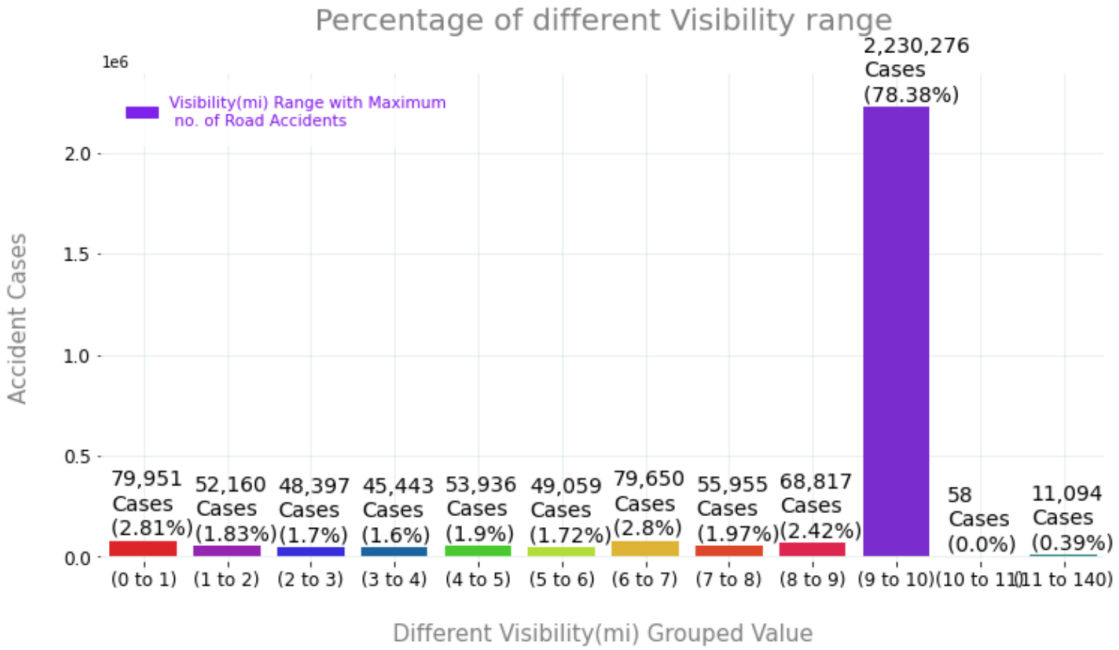


Figure 2.24: Visibility Ranges at Time of Incident

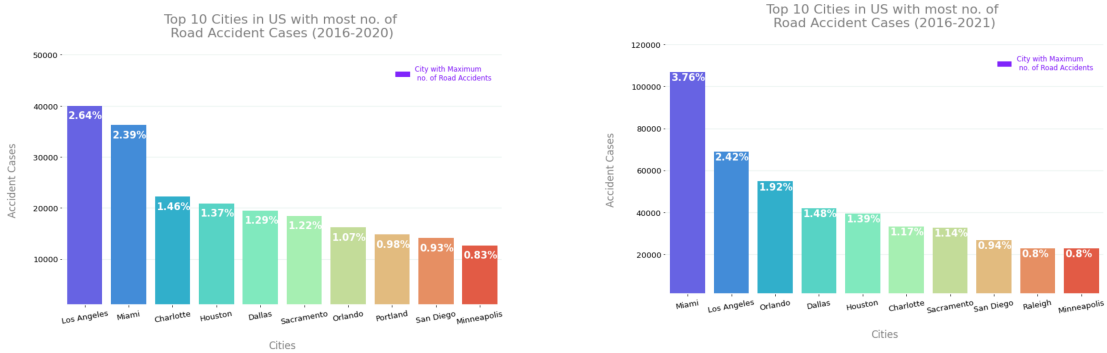
Over 2 million cases (78.38%) had a visibility range of 9 to 10 miles meaning, in the majority of this cases visibility was most likely not a factor. In 2.81% of cases visibility was only 0 to 1 miles, with this knowledge we can assume visibility was a main factor in the 2.81% of cases that an accident occurred.

Chapter 3 Contribution Shifts & Data Analysis Focusing in on the Year 2021

Comparing and contrasting the already discovered data on vehicle accidents up to 2020 and the newly found data we were able to pull up to 2021 will give good insight on changes for the current year, 2021 and why these changes may have occurred.

3.1 Location Changes for 2021

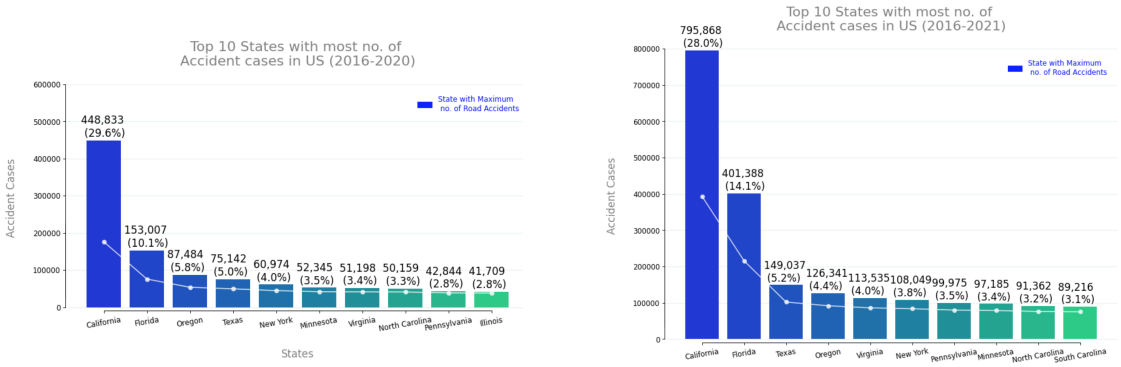
As we have previously mentioned, looking at Figure 3.1, the city with the highest amount of accidents on record took a shift from Los Angeles by the end of 2020 to Miami by the end of 2021. Now we will investigate why this is the case: According to macrotrends website Miami's population received a 0.74% increase from 2020 while Los Angeles had only 0.1% increase in population from 2020 to 2021 [17, 16]. Based on the US Census Bureau Data and internal metrics from 2020, Florida ranks the number one destination for American's looking to relocate to a new state. According to Movement Mortgage, because of Florida's low cost of living, appealing job market, and zero income tax policy Florida looks the most appealing to the majority of American



(a) (2016-2020) (b) (2016-2021)

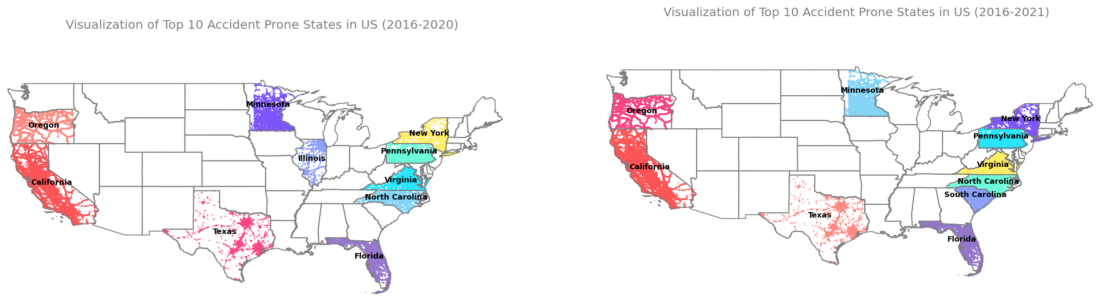
Figure 3.1: Top 10 Cities with Most Recorded Accidents Comparison Analysis

citizens more than it has ever been [3].



(a) (2016-2020) (b) (2016-2021)

Figure 3.2: States with Most Recorded Accidents Comparison Analysis

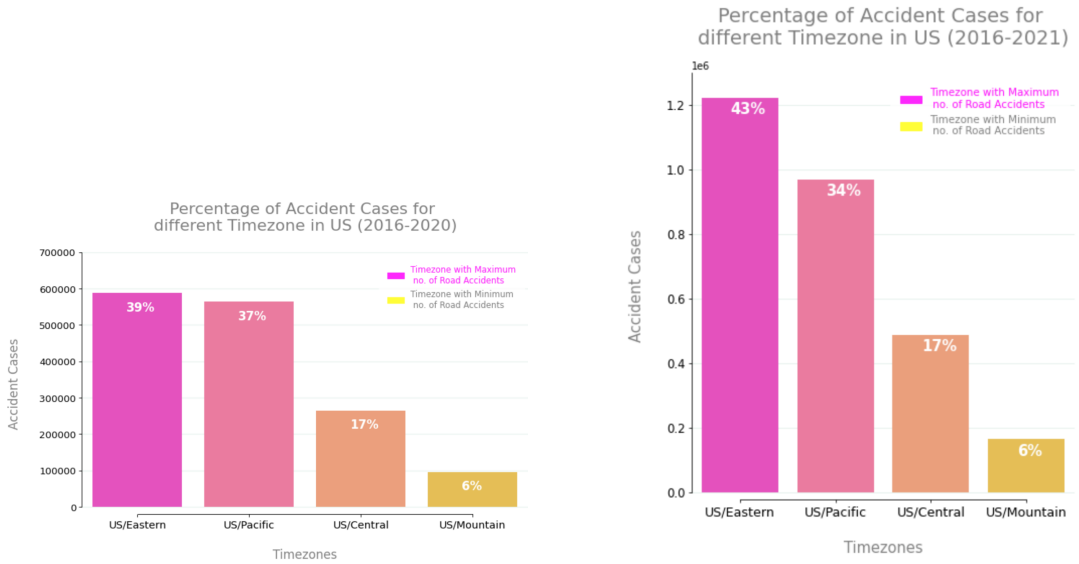


(a) (2016-2020) (b) (2016-2021)

Figure 3.3: Visualization of Top 10 States

From the comparison graphs in Figure 3.2 we can see that California still remains the state with the highest amount of car accidents, but with the incidents nearly doubling from 2020 to 2021. While it would be likely for a person to assume that this could be caused from population rising in California, the facts are that California's population declined from 2020 to 2021 by 0.66% [16]. According to Los Angeles's Department of Transportation one of the number one causes of vehicle accidents in 2021 was distracted driving. Data taken shortly after the end of 2021 showed around 80% of people driving in California are actively using technology [4].

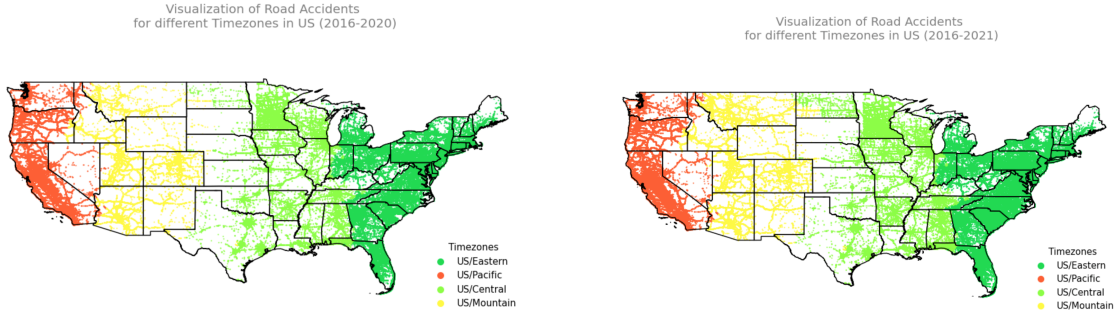
Figure 3.4 shows that US Eastern Time remains the highest percentage from the end of 2020 to the end of 2021. The percentage of the amount of accidents along the eastern side slightly increases from 39% to 43%. The eastern timezone contains Florida, New York, and a couple other high city populated states which we can see on our list of top 10 sates above in Figure 3.2. Since the amount of accidents increased



(a) (2016-2020) (b) (2016-2021)

Figure 3.4: Timezone Comparison Analysis

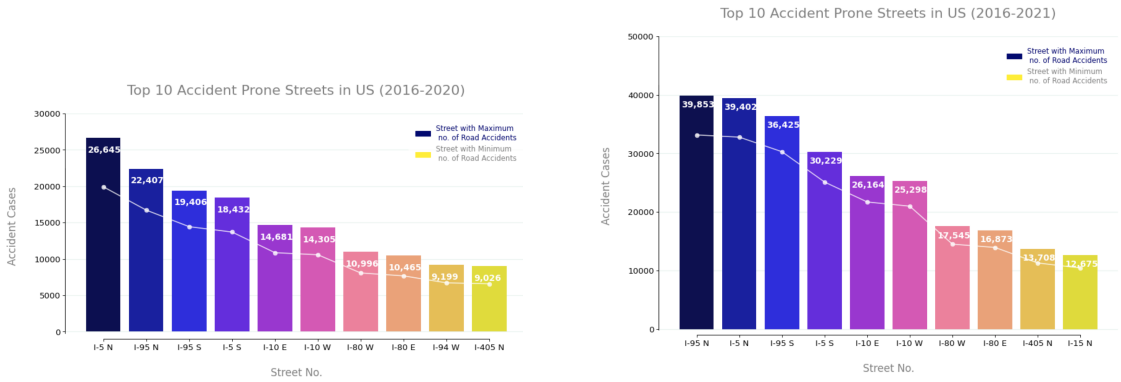
significantly in Florida we can concur that the amount of accidents in the eastern timezone would increase as well (since Florida belongs to the eastern timezone).



(a) (2016-2020) (b) (2016-2021)

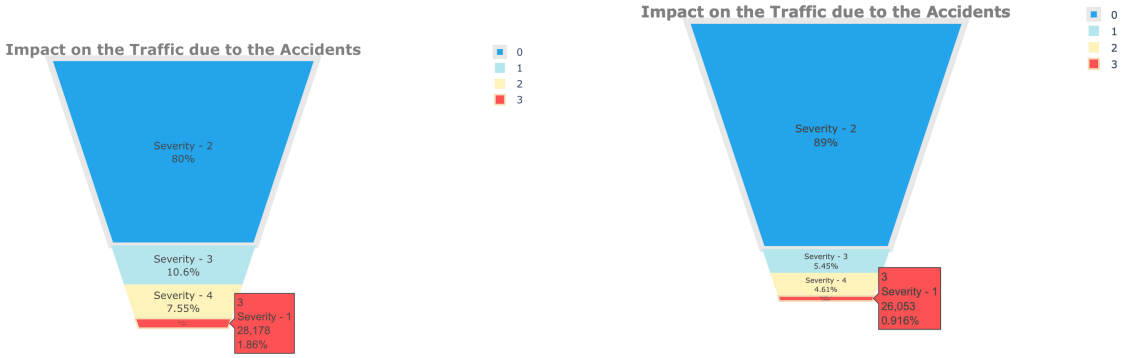
Figure 3.5: Timezone Map Visualization

From 2020 to 2021 the highest accident highway went from I-5 N to I-95 N. I-95 N had a total of 13,208 accidents in the year 2021. I-95 is an interstate that runs from Florida all the way up to Maine making it a popular choice for those driving along the east coast. A combination of I-95 N's heavy congestion, work zones, and high speed limit (60 to 70mph) causes major hazards for drivers [2].



(a) (2016-2020) (b) (2016-2021)

Figure 3.6: Top Interstate's Prone to Accidents Comparison Analysis

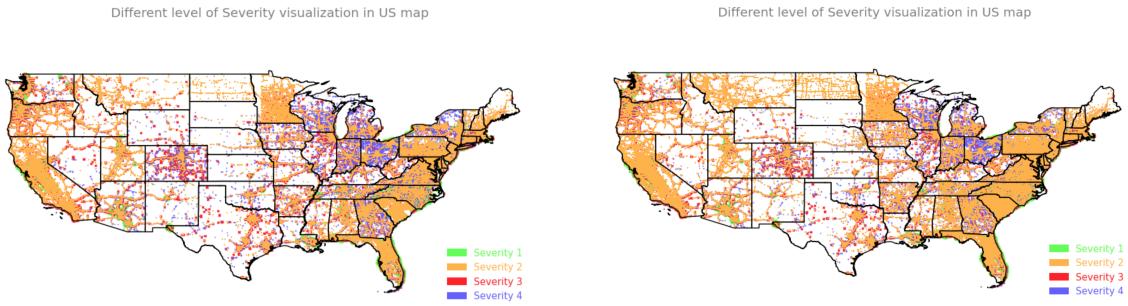


(a) (2016-2020) (b) (2016-2021)

Figure 3.7: Severity Comparison Analysis

3.2 Levels of Severity for 2021

The graphs above compare severity on a scale from one to four where one indicates the most minimal effect on traffic (short delays, no inconvenience, etc.) and four indicates the most significant impact on traffic (long delays). As we can see from 2020 to 2021 level 2 severity slightly increased while all other severity's slightly decreased. We can concur that there were more accidents in 2021 that had light to moderate impact on traffic. As we can see from the Severity Map below the change in severity accidents around the entire United States still remains minimal with a slight increase in states like, Montana and North Dakota while other states see little to no change.



(a) (2016-2020) (b) (2016-2021)

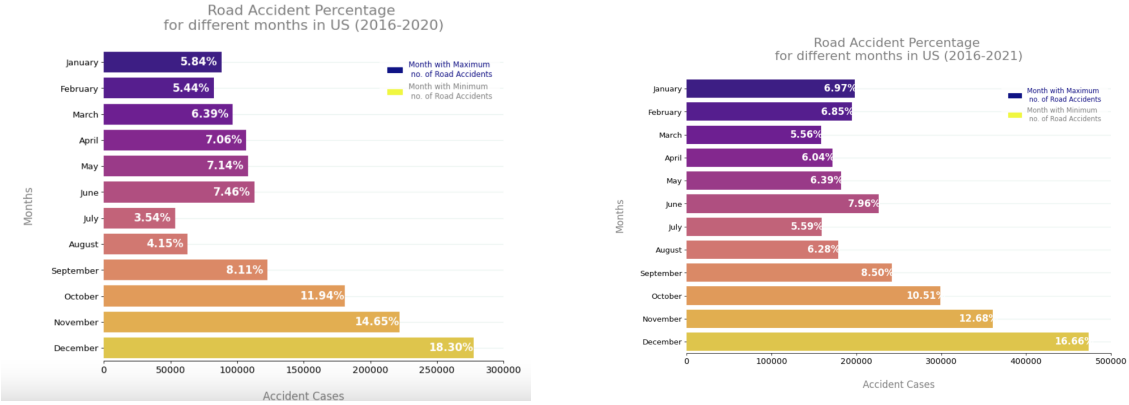
Figure 3.8: Severity Map Visualization

3.3 Time of Incident Analysis for 2021

(a) (2016-2020) (b) (2016-2021)

Figure 3.9: Accident Duration Comparison Analysis

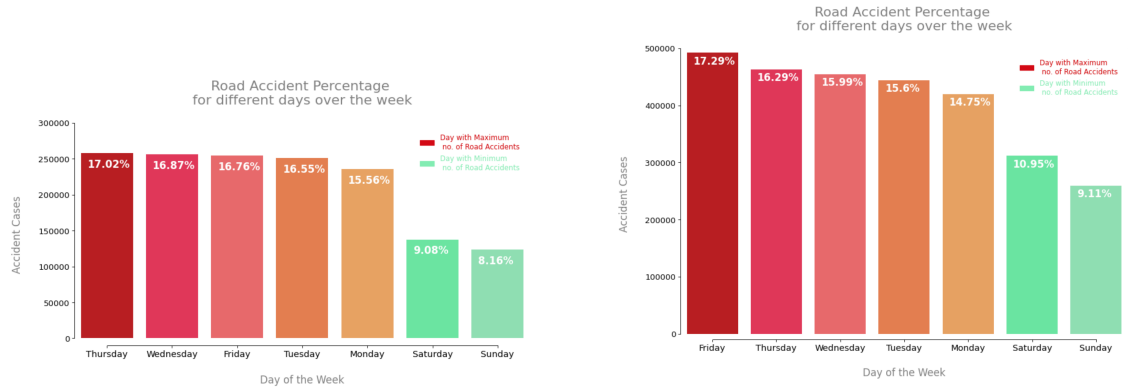
The graphs above show the differences in the amount of time traffic was impacted from the result of an accident. We can see that from 2020 to 2021 the amount of time impacted most often remains the same at 6 hours. Six hours is the highest amount of time recorded in the database to affect traffic and remains the most overall time that does occur when an accident impacts traffic. The amount of accidents causing 6 hour traffic impacts did decrease by half from 2020 to 2021. According to the 2021 Global Traffic Scorecard, Americans lost 3.4 billion hours due to congestion in 2021 which may seem like a lot, but this is actually down by 42% from pre-covid times. According to Bob Pishue, a transportation analyst at INRIX, COVID-19's impact on traffic has continued up to 2021 and although congestion has increased 28% in the year 2021 there remains notable changes to commuting during the pandemic such as reduced travel times volumes of vehicles on the road and fewer downtown traffic [6].



(a) (2016-2020) (b) (2016-2021)

Figure 3.10: Months Comparison Change to 2021

Looking at the percentage of accidents that have occurred each month from the end of 2020 to the end of 2021, obviously the amount of accidents in every month increased from 2020 to 2021 but the percentage of accidents that occurred each month definitely differs between the two graphs. December remains the month with the most amount of accidents while the month with the least amount of accidents changes from July in 2020 to March in 2021. The year 2020 saw fewer car accidents overall in the months March, April, May and June. The most likely cause is that during this time, it was the height of the COVID-19 shutdown. Looking at the graphs, we can see this causes an impact on the year 2021's data having less overall accidents in the month of March from 2016 to 2021 which is what the graph above, in part b, depicts.

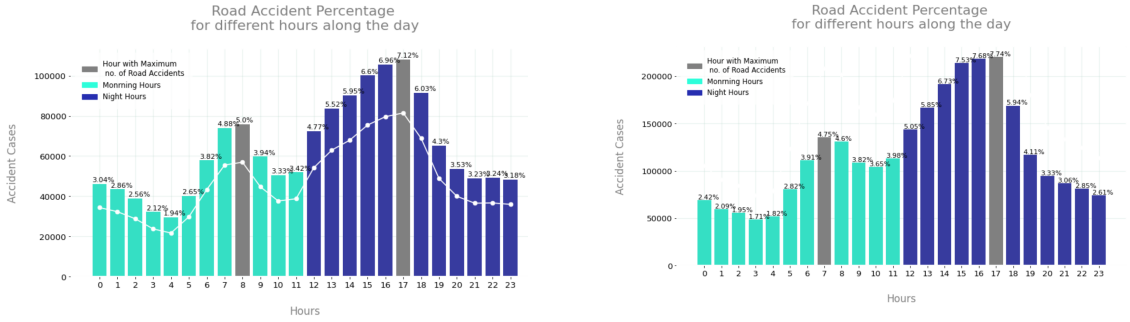


(a) (2016-2020) (b) (2016-2021)

Figure 3.11: Days in a Week Comparison Analysis

As we can see from Figure 3.11, the most amount of accidents remain during the "work-week" Monday through Friday while the most amount of accidents in one day

overall changed from Thursday to Friday in 2021. The change is only 1% higher accidents than that of Thursday. It makes sense that accidents are more common during the work-week because more people have a responsibility to drive to and from work where weekend activities remain a choice on whether or not you decide to drive.



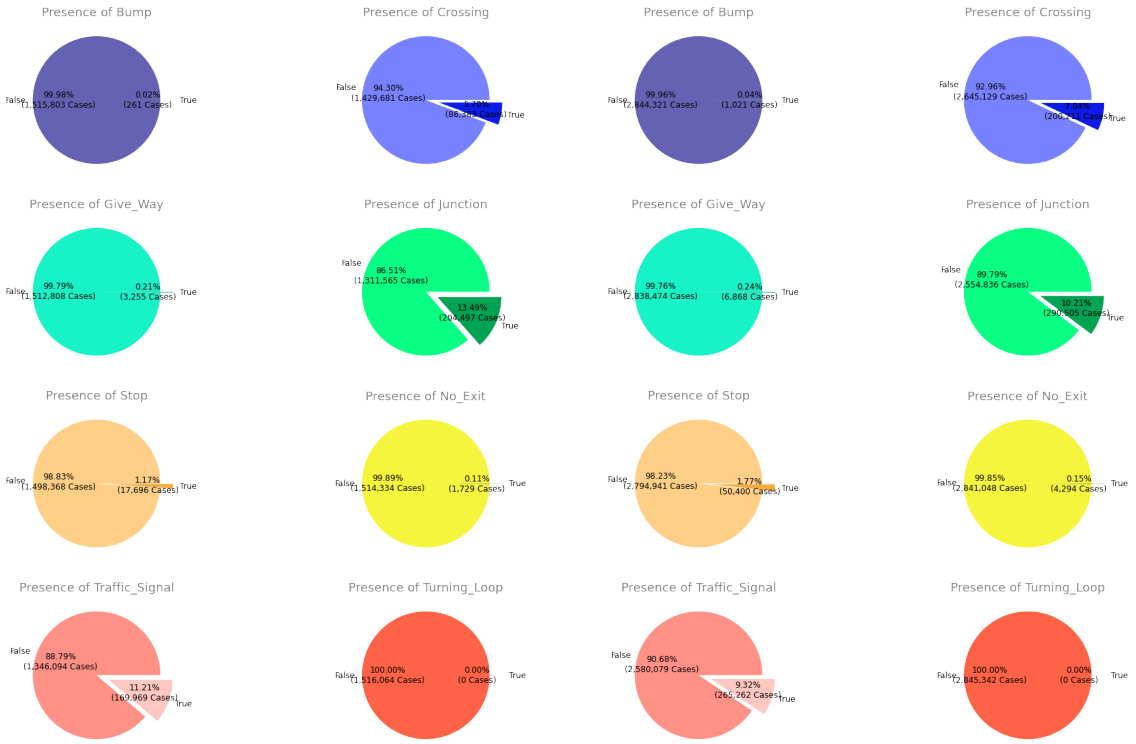
(a) (2016-2020) (b) (2016-2021)

Figure 3.12: Hours in a Day Comparison Analysis

Previously mentioned in chapter 2, the deadliest morning hour changed in 2021 from 8:00AM to 7:00AM and the deadliest night hour remains at 5:00PM. Although not much data has been recorded about the American population work week schedules, this data recorded on vehicle accidents could indicate that people are starting work earlier than in the past.

3.4 Road Condition Comparison from the end of 2020 up to the end of 2021

Looking at the pie charts in Figure 3.13 we can see that there were no dramatic changes from 2020 to 2021. The changes that did occur going from 2020 to 2021 were very slight. The presence of a bump, crossing, give way, stop sign and no exit sign increased slightly meaning that accidents with those settings did occur in the year 2021 while the presence of a junction and traffic signal decreased in the year 2021 meaning that accidents with that present either did not occur at all or occurred slightly. Accidents with the presence of a turning loop have either not occurred or they seldom did, this would mean that they have not been submitted into a database

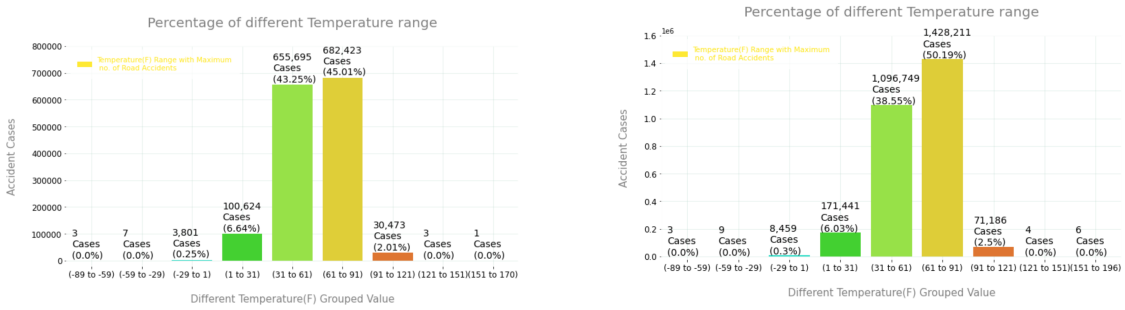


(a) (2016-2020) (b) (2016-2021)

Figure 3.13: Road Condition Comparison Analysis

since 2016 up to 2021. For the accident to have not been submitted into a data base would entail a police officer not being present at the scene and the person that this occurred to would have had to manage the situation on their own.

3.5 Weather Comparison Analysis



(a) (2016-2020) (b) (2016-2021)

Figure 3.14: Temperature Comparison Analysis

The changes in temperature regarding the percentage of accidents that occurred

changed very slightly but if we look at the amount of accidents that occurred in the year 2021 accidents that happened when the temperature ranged from 61 to 91 degrees Fahrenheit more than doubled. In the year 2021 the average temperature was 61 to 91 in the months of June, July, August, and September. Although the months that fell into the most common temperature range are only 4 out of 12, the other 6 contain more drastic temperatures at times that would equate to the other 50% of temperatures a bit more spread out with accidents occurring in the range 31 to 61 degrees Fahrenheit still having around 39% of the accidents that happened. Below is a line graph of average temperatures of each month in the United States from January 2019 up to May 2022 in degrees Fahrenheit so we can have a better since of what the temperature was around, when these accidents occurred [7].

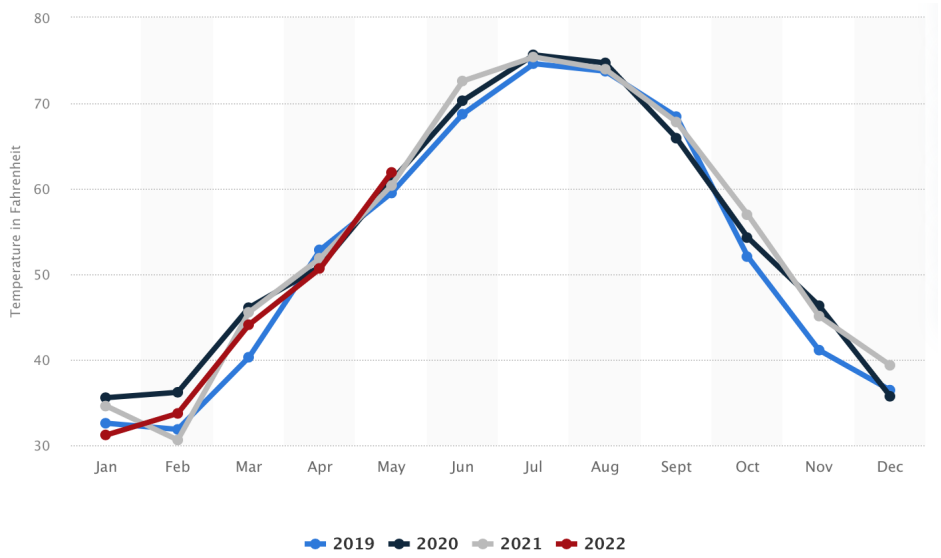
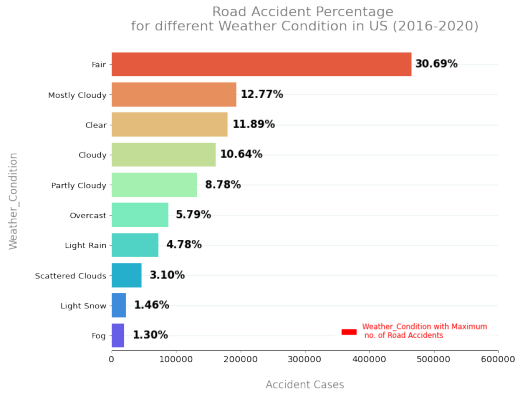


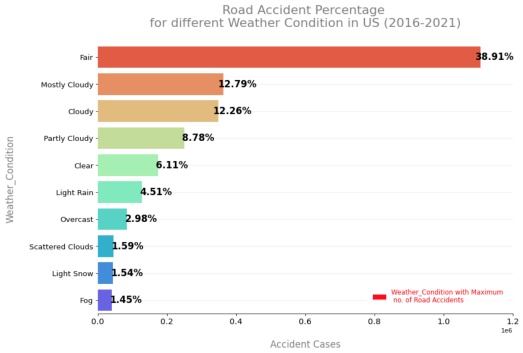
Figure 3.15: Average Monthly Temperature in the U.S. (2019-2022)

In 2021, the amount of accidents that happened when the weather was "fair" increased by more than 8%. This means that overall, American's are getting into slightly less accidents when there are "poor" weather conditions such as light rain, overcast, and scattered clouds.

The trend of accident percentages during different humidity ranges stays close to the same from the end of 2020 to the end of 2021. Although we can visually see that each bar in graph "a" increases dramatically more than in graph "b" this is to be expected because of the high amount of accidents that occurred in the year 2021.

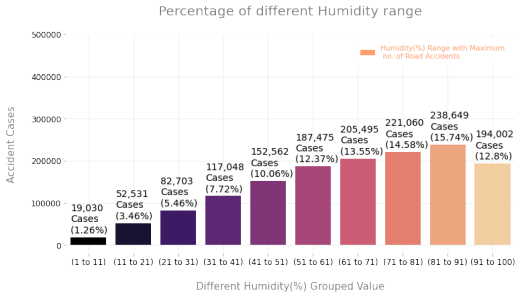


(a) (2016-2020)

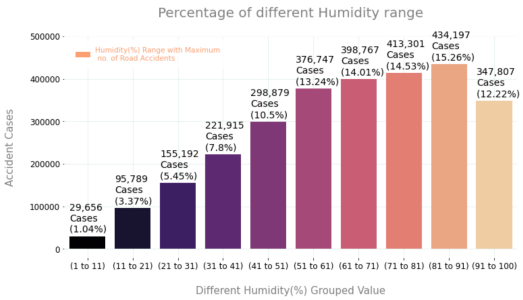


(b) (2016-2021)

Figure 3.16: Types of Weather Comparison Analysis from (2016-2020) to (2016-2021)



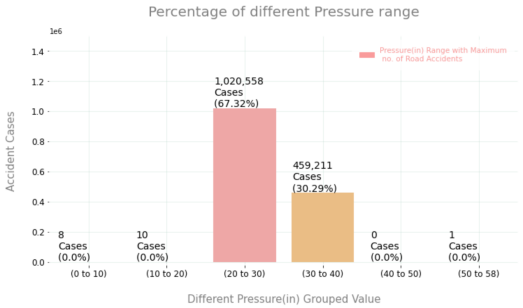
(a) (2016-2020)



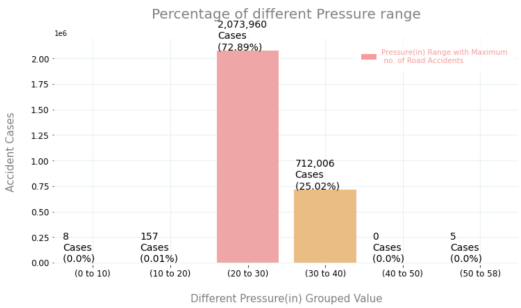
(b) (2016-2021)

Figure 3.17: Humidity Comparison Analysis

Different pressure percentages from 2016 to 2021 remains similar. The noticeable



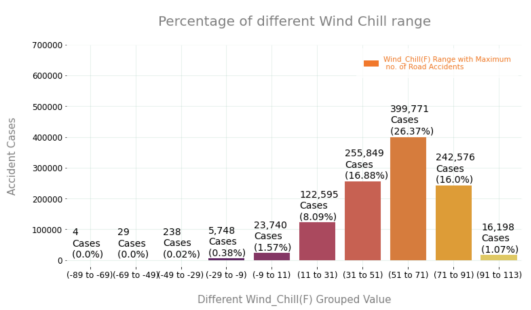
(a) (2016-2020)



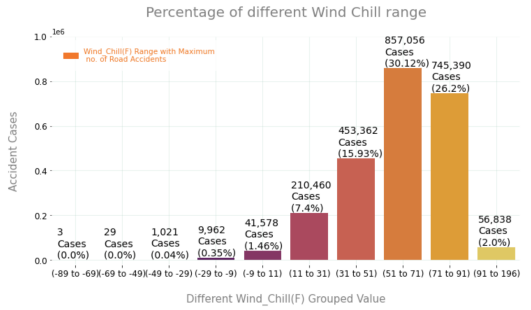
(b) (2016-2021)

Figure 3.18: Pressure Comparison Analysis

accident changes for the year 2021 are when the air pressure ranges from 10 to 20 inches. Up to 2020 there are only 10 accidents that have happened during the 10 to 20 range while in 2021 there are 157 accidents that have occurred since 2016. This means that in the year 2021 there was a 1,470% increase of accidents that occurred within the 10 to 20 pressure range.



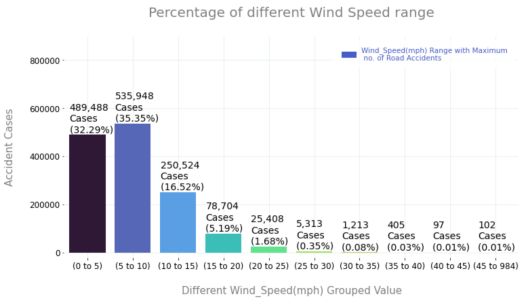
(a) (2016-2020)



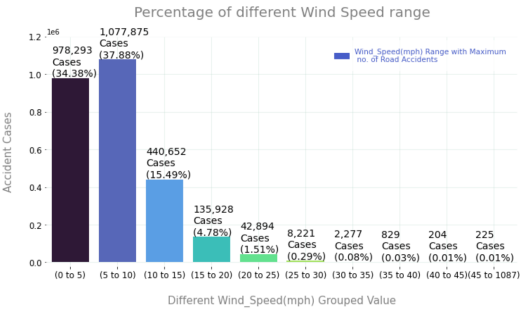
(b) (2016-2021)

Figure 3.19: Wind Chill Comparison Analysis

As we can see looking at the Wind Chill Figure, the accidents that happened the most during 2021 remain in the wind chill range of 51 to 71 degrees Fahrenheit and the others fall accordingly all the way down to the -89 to -69 degrees Fahrenheit range containing still the least amount of accidents happening within that range.



(a) (2016-2020)

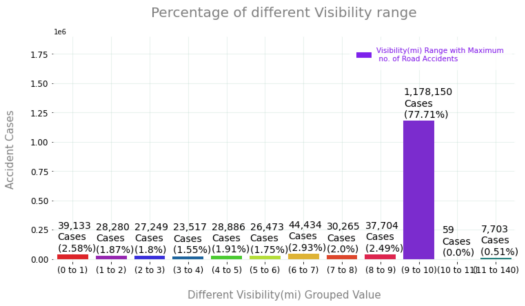


(b) (2016-2021)

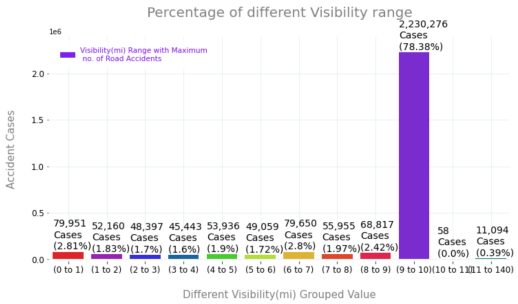
Figure 3.20: Wind Speed Comparison Analysis

The Wind Speed Comparison Analysis is similar to the Wind Chill Comparison Analysis in the sense that for the year 2021 the highest cases remain the same while the others fall in accordance with the data from 2016 to 2020.

Below, we can see the few changes in Visibility from 2016-2020 to 2016-2021.



(a) (2016-2020)



(b) (2016-2021)

Figure 3.21: Visibility Comparison Analysis

Chapter 4 COVID-19 Effects and Expectations

4.1 Introduction

Coronavirus swept through the nation approximately in March of 2020. During this time many people were urged to go into lockdown including several workers who were able to work remotely, all public school systems, and most non-essential workers.

4.2 Impact on Truck Drivers

Occupations that were deemed "essential" kept up and running during unprecedented times. With a huge majority of the population working from home, people were also encouraged to stay at home for other means than just working, one of those means being shopping. Online shopping increased significantly during the pandemic and commercial truck drivers, which remained essential, worked overtime during the pandemic, to try and meet the demand. According to the National Safety Council, this lead to more dangerously "drowsy" drivers who became prone to causing truck accidents with injuries on the road. Despite there being a reduced amount of drivers traversing on the U.S. roadways, statistics showed an increase in fatal accidents in the beginning months of 2020. The global pandemic has left it's mark on the commercial trucking industry in numerous ways. With an increase in demand from consumers comes with an increase in demand from employers. After the pandemic shutdown hit, the Federal Motor Carrier Safety Administration temporarily waved the working hour limits for commercial truck drivers. When drivers become fatigued from lack of sleep and being overworked that can directly lead to a collision. Although truck drivers hours are no longer waived, there are still commercial truck drivers working

overtime to meet the demand of the steadily increased online market [5]. The data below tells although there were less trucks on the U.S. roadway overall the ones that were had accidents enough to only decrease from 2019 by 10% while still increasing overall by 62% from 2008 [12].

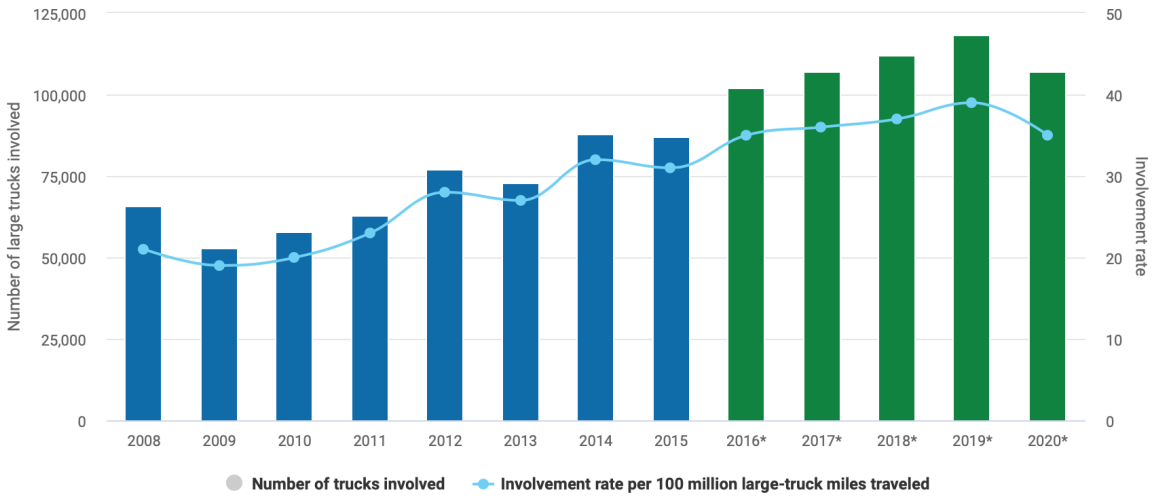


Figure 4.1: Large Truck Involvement in Injury Crashes (2008-2020)

The year 2021, still dealing with the aftermath of COVID-19, saw a 13% increase in the number of deaths in accidents involving a large truck compared to the year 2020. The National Highway Traffic Safety Administration (NHTSA) released the latest data for crashes involving a commercial truck for the year 2021. In order for the NHTSA to collect the correct information on the trucking industry, they considered any commercial truck weighing at or above 10,000 pounds to be considered a "large truck" for data collection purposes. The alarming increase in trucking accidents over the past years have called for trucking safety groups to advocate for the federal administration to quickly approve a number of pending regulations. One of the regulations involves mandatory installation of speed limiters on all commercial trucks [9].

4.3 Impact on Driving Behaviors and Crash Severity

A study done by Dong, Xie, and Yang [1] after the COVID-19 pandemic hit showed insights into the effect of driving behaviors on safety during the event of the pandemic.

The group predicted that risky driving behaviors, such as speeding and failing to signal were taking place more often after March of 2020, when lockdown went into effect in many places, which would have resulted in higher rates of severe crashes. To test their hypothesis, the group used structural equation modeling (SEM) for multigroups to capture the complex interrelationships between crash injury severity, the context of COVID-19, driving behaviors and other risk factors for two groups, being highways and non-highways. The SEM constructs two latent variables, aggressiveness and inattentiveness. These are indicated by risk driving behaviors such as speeding, drunk driving, and distraction. SEM is expressed as:

$$y_i = x_i\beta + \gamma_1 Args_i + \gamma_2 Intv_i + \varepsilon_i^y$$

$$y_i = 1, if y_i^* > \varphi, y_i = 0, otherwise$$

$$Agrs_i = \alpha_1 COVID_i + \varepsilon_i^{Args}$$

$$Intv_i = \alpha_2 COVID_i + \varepsilon_i^{Intv}$$

where i (i=1,2,3,...,N) is the index for crashes; y_i^* is propensity of crash severity (the larger the more likely to be involved with severe crashes); y_i is the observed injury severity (0 for non-severe injury, 1 for severe injury) for crash i; x_i is a vector of observed variables that indicate collision type, road and environmental features for crash i; β is a vector of coefficients corresponding to x_i; $Args_i$ is a latent variable indicated by aggressive driving behaviors in crash i; $Intv_i$ is a latent variable indicated by inattentive driving behaviors in crash i; $COVID_i$ is an observed variable that indicates the crash i is during the presence of COVID-19 pandemic; γ_1 and γ_2 are the coefficients of the two latent variables; α_1 and α_2 are the coefficients of $COVID_i$; ε_i^y, ε_i^{Agrs}, and ε_i^{Intv} are normally distributed error terms; and φ is a threshold to determine crash severity outcome. In multigroup SEM with equal thresholds, φ is held the same across highway and non-highway groups, while other parameters are not contained. Similarly, in multigroup SEM with equal regressions, β, α_1, α_2, γ_1, and γ_2 are held the same across groups, while other parameters are not constrained.

For multigroup SEM with no constraint, all parameters are freely estimated across groups. The measurement model for SEM is formulated as:

$$DB^{Agrs} = Agrs\Lambda^{Agrs} + \delta^{Agrs}$$

$$DB^{Intv} = Intv\Lambda^{Intv} + \delta^{Intv}$$

where DB^{Agrs} is a (N x p) matrix of the observed driving behaviors associated with aggressiveness, and DB^{Intv} is a (N x q) matrix of the observed driving behaviors associated with inattentiveness; Agrs is a (N x 1) vector of the latent variable aggressiveness, while Intv is a (N x 1) vector of the latent variable inattentiveness; Λ^{Agrs} is a (1 x p) vector of factor loading for aggressiveness, and Λ^{Intv} is a (1 x q) vector of factor loading for inattentiveness; δ^{Agrs} and δ^{Agrs} are (N x p) matrices of gaussian errors.

After extracting the data from the SEM models, results suggest aggressiveness and inattentiveness of drivers increased significantly after the outbreak. This leads to a higher likelihood of severe crashes. This study provides insights into the effect of changing driving behaviors on safety during an event such as COVID-19 [1].

4.4 Fatalities

The National Safety council revealed, despite there being an overall reduced amount of drivers on the road, statistics showing an increase in fatal accidents by 14% when comparing the months March of 2019 to March 2020. Comparing rates of each state, Tennessee is one of the top contributing states to this statistic with a 6% increase in traffic deaths within the first few months of the year 2020. Similar to the study done by Dong, Xie, and Yang, the National Safety Council attributes the spike in roadway deaths to drivers' tendencies to drive more recklessly with fewer cars on the road [12]. In the year 2021, the Department of Transportation National Highway Administration reported traffic fatalities increased by 18.4% from the year 2020. An estimated 20,160 vehicle fatalities occurred during the first half of 2021 which has been

the highest since 2006. It is important to note, because of these recent statistics, in November 2021, the Infrastructure Investment and Jobs Act was signed into law. The bill includes $5 billion for a new program to support local initiatives to prevent death and serious injuries on roads and streets, and approximately $39 billion in new funding to modernize the nation's public transit system [15].

Chapter 5 Conclusion and Recommendations

5.1 Inspiration

This thesis was inspired by recent work by Satyabrata Roy who explored factors that influence US road accidents from 2016 up to 2020. Roy also used A Countrywide Traffic Accident Dataset, called "US Accidents" [11]. Studying the given data is useful for purposes such as real-time car accident prediction, accident hotspot locations, causality analysis, the impact of environmental stimuli, and the effect of COVID-19 on traffic behavior and accidents.

5.2 Recommendations and Future Research

Before explaining the process of using and applying machine learning to future research it is paramount to understand what machine learning is and why it is important. Machine learning is best described as a subset of artificial intelligence technologies, it involves training a machine to learn more quickly and intelligently. Machine learning is an optimization process for AI technologies with it being responsible for providing better and faster training to AI solutions. Machine learning has existed for years, but machine learning processes have recently taken prominence due to several technological improvements in recent years such as wider access to large volumes and varieties of data, i.e. "big data". It is also prominent because of increasing processing power which allows AI applications to complete calculations quicker than ever before [18].

For the conclusion of this study, we will present some steps for data analysis using machine learning techniques such as K-Nearest Neighbors, Decision Tree, Random

Forest and Logistic Regression. First, we import the libraries needed for each technique. The necessary libraries defined for our machine learning are:

1. **KNeighborsClassifier (from sklearn.neighbors):** Implements classification based on voting by nearest k-neighbors of target point, t.
2. **DecisionTreeClassifier (from sklearn.tree):** A class capable of performing multi-class classification on a dataset.
3. **RandomForestClassifier (from sklearn.ensemble):** A classification algorithm consisting of many decision trees. Uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.
4. **LogisticRegression (from sklearn.linear_model):** Predicts the probability of a categorical dependent variable.
5. **train_test_split (from sklearn.model_selection):** Splits data arrays into two subsets- for training data and testing data.
6. **GridSearchCV (from sklearn.model_selection):** Performs hyperparameter tuning in order to determine the optimal values for a given model.
7. **SelectFromModel (from sklearn.feature_selection):** A meta-estimator that determines the weight importance by comparing to the given threshold value.
8. **classification_report (from sklearn.metrics):** Used to measure the quality of predictions from a classification algorithm.
9. **confusion_matrix (from sklearn.metrics):** Evaluates the accuracy of a classification.
10. **accuracy_score (from sklearn.metrics):** Computes the accuracy, either the fraction (default) or the count (normalize=False) of correct predictions.
11. **roc_curve, auc (from sklearn.metrics):** Presents a graph showing the performance of a classification model at all classification thresholds. The curve plots two parameters, true positive rate and false positive rate.

Next, we import the dataset we have been using throughout our research called "US Accidents". Then, we get each variable information using df.info(). Next, we can use "Start_Time" and "End_Time" to extract year, month, day, hour, weekday, and time to sort out accident periods. Once we have this data, we want to drop the outliers which are rows with negative time duration. We can replace the existing outliers with median values. After these steps we are able to check variables related to time and road accidents such as the maximum time it took to clear a road accident which

is 20 days and minimum time to clear a road accident which is 2 days.

Early research shows the following features being most important based on the changing variables discovered using machine learning. We select these variables, drop rows with missing values and select a point of intersection so the data does not become to large to handle. In our case, we choose the location point of intersection as North Carolina. One example of information we can represent using the solo state is a pin point of accidents by county from 2016 to 2021 which is represented below.

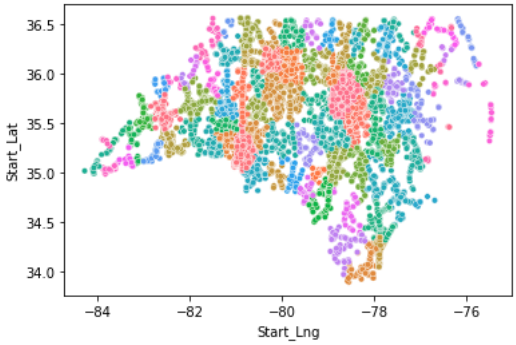


Figure 5.1: Map of Accidents in NC by County (2016-2021)

Once we have all our data we need into the program we can finally predict accident severity using various supervised machine learning algorithms. We prepare the data by using our library train_test_split which we had previously imported. Using the technique of K-Nearest Neighbors (KNN with 6 neighbors) algorithm we get a knn.score and accuracy_score of 89.4%. We can also use the K-Nearest Neighbors algorithm to generate a plot that shows the accuracy for each number of neighbors to guide the optimization.

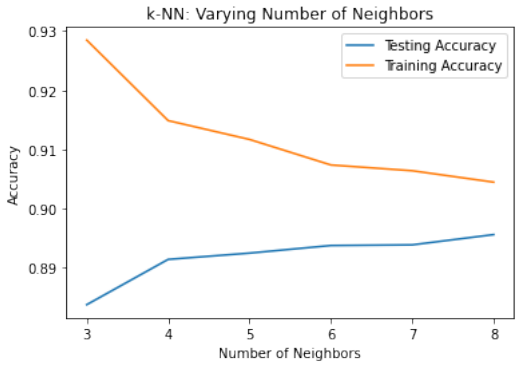


Figure 5.2: Accuracy Compared to the Number of Neighbors

The plot shows which number of neighbors is most optimal for the best result in

accuracy. Using the Random Forest algorithm we are able to get an accuracy result of 90.9% which is a better outcome than KNN with 6 neighbors. We can also look at the importance score for each feature. In this case, using Random Forest technique to visualize the most important features, we can see below that distance is the most important feature.

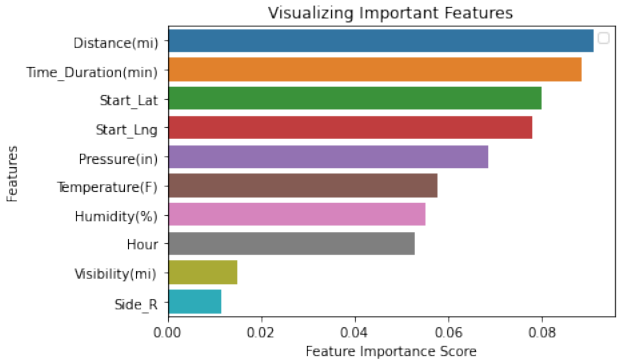


Figure 5.3: Visualization of Important Features

The previous plot indicates that distance travelled and time_duration are leading factors of accident severity.

Early results indicate that, Random Forest (instead of K-Nearest Neighbors) algorithm is a better in predicting with greater precision, accident severity. Further research can use a combination of machine learning algorithms to further establish leading factors in US road accidents and their levels of severity.

Bibliography

[1] Dong, X., Xie, K., & Yang, H. (2022). How did COVID-19 impact driving behaviors and crash Severity? A multigroup structural equation modeling. Accident; analysis and prevention, 172, 106687. https://doi.org/10.1016/j.aap.2022.106687

[2] Evans, L., &; Rodney, T. B. (2022, February 16). I-95: Car accidents on the East Coast's main highway. JD Supra. Retrieved October 7, 2022, from https://www.jdsupra.com/legalnews/i-95-car-accidents-on-the-east-coast-s-6901946/

[3] Ganz, P. (2022, May 19). How many people are moving to Florida every year? Make Florida Your Home. Retrieved October 7, 2022, from https://www.makefloridayourhome.com/blog/how-many-people-are-moving-to-florida-every-year-

[4] Gonzalez, S. (2022, February 4). Traffic deaths in California are on the rise. here's how LA and other big cities are trying to change that. KQED. Retrieved October 7, 2022, from https://www.kqed.org/news/11903812/traffic-deaths-in-california-are-on-the-rise-heres-how-la-and-other-big-cities-are-trying-to-change-that

[5] How Has Covid-19 Affected the Trucking Industry and Driver Safety? Warren and Griffin. (2022, May 25). Retrieved October 12, 2022, from https://www.warrenandgriffin.com/how-has-covid-19-affected-the-trucking-industry-and-driver-safety/

[6] Inrix. (2021, December 7). Americans lost 3.4 billion hours due to congestion in 2021, 42% below pre-COVID. Inrix. Retrieved October 7, 2022, from https://inrix.com/press-releases/2021-traffic-scorecard/

[7] Jaganmohan, M. (2022, July 5). Monthly average Temperature United States 2022. Statista. Retrieved October 7, 2022, from https://www.statista.com/statistics/513628/monthly-average-temperature-in-the-us-fahrenheit/#statisticContainer

[8] Kademenos, V. (2020, October 13). Most common causes of car accidents in Ohio. KWHDW. Retrieved October 7, 2022, from https://www.ohattorneys.com/common-causes-car-accidents-ohio/

[9] Katz, R. N. (2022, June 4). Feds Record 13% Increase in Trucking Accident Deaths in 2021. Georgia Injury Law Blog. Retrieved October 17, 2022, from https://www.georgiainjurylawblog.com/feds-record-13-increase-in-trucking-accident-deaths-in-2021/

[10] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (2021).

[11] 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, 2021.

[12] National Safety Council. (2022, April 6). Large trucks. Injury Facts. Retrieved October 12, 2022, from https://injuryfacts.nsc.org/motor-vehicle/road-users/large-trucks/

[13] The majority of car accidents happen close to home. Law Office Of Deborah M. Truscello. (2021, March 3). Retrieved October 7, 2022, from https://www.truscellolaw.com/blog/2021/03/car-accidents-happen-home/

[14] U.S. Department of Transportation Federal Highway Administration. (2022, July 26). How do weather events impact roads? How Do Weather Events Impact Roads? - FHWA Road Weather Management. Retrieved October 7, 2022, from https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

[15] U.S. Government Accountability Office. (2022, January 25). During COVID-19, Road Fatalities Increased and Transit Ridership Dipped. U.S. GAO. Retrieved October 17, 2022, from https://www.gao.gov/blog/during-covid-19-road-fatalities-increased-and-transit-ridership-dipped

[16] United Nations- World Population Prospects. (2022). Los Angeles Metro Area Population 1950-2022. MacroTrends. Retrieved October 7, 2022, from https://www.macrotrends.net/cities/23052/los-angeles/population

[17] United Nations- World Population Prospects. (2022). Miami metro area population 1950-2022. MacroTrends. Retrieved October 7, 2022, from https://www.macrotrends.net/cities/23064/miami/population

[18] Why is Machine Learning Important? Colorado State University Global. (2021, July 6). Retrieved November 6, 2022, from https://csuglobal.edu/blog/why-machine-learning-important

Appendix

5.3 All Necessary Imported Libraries Defined

1. **numpy as np:** Performs a wide variety of mathematical operations on arrays.
2. **pandas as pd:** Provides fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive.
3. **matplotlib.pyplot as plt:** Able to utilize various plots such as Line Plot, Histogram, Scatter, 3D Plot, Image, Contour, and Polar.
4. **matplotlib.ticker as ticker:** Sets a tick for every integer multiple of a base within the view interval.
5. **matplotlib.patches as mpatches:** An object you can add to the plot which are customizable in all the typical Matplotlib ways.
6. **seaborn as sns:** Provides a high-level interface for drawing attractive and informative statistical graphics.
7. **calendar:** Allows you to perform date, month, and calendar-related operations while even letting you manipulate your code for some specific day or month of the year.
8. **ploty as pt:** Creates interactive, publication-quality graphs.
9. **graph_objs as go (from ploty):** Contains descriptions of each valid property as Python docstrings.
10. **ploty.express as px:** Operates on a variety of types of data and produces easy-to-style figures.
11. **ploty.figure_factory as ff:** Creates very specific types of plots that were at the time of their creation difficult to create with graph objects and prior to the existence of Plotly Express.
12. **matplotlib.patheffects as PathEffects:** Provides functionality to apply a multiple draw stage to any Artist which can be rendered via a path.
13. **descartes:** Provides a nicer integration of Shapely geometry objects with Matplotlib.
14. **geopandas as gpd:** Extends the datatypes used by pandas to allow spatial operations on geometric types.
15. **distance (from Levenshtein):** A text similarity measure that compares two words and returns a numeric value representing the distance between them.

16. **product (from itertools):** Returns the cartesian product of the provided iterable with itself for the number of times specified by the optional keyword "repeat".

17. **fuzz (from fuzzywuzzy):** Implements application level checks to catch application/ logical bugs.

18. **process (from fuzzywuzzy):** Finds the best matches in a list or dictionary of choices, returns a list of tuples containing the match and it's score. If a dictionary is used, it also returns the key for each match.

19. **pdist, squareform (from scipy.spatial.distance):** Pairwise distances between observations in n-dimensional space. Coverts a vector-form distance vector to a square-form distance matrix, and vice-versa.

20. **Point, Polygon (from shapely.geometry):** Functions that check if a point is within a polygon and checks if a polygon contains a point.

21. **geoplot:** Provides a selection of easy-to-use geospatial visualizations.

22. **Nominatim (from geopy.geocoders):** A tool to search OpenStreetMap data by address or location (geocoding).

23. **warnings:** Alerts the user of some condition in a program, where that condition (normally) doesn't warrant raising an exception and terminating the program.

5.4 Partial Python Codes

```
# import all necesary libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.ticker as ticker
import matplotlib.patches as mpatches
%matplotlib inline
import seaborn as sns
import calendar
import plotly as pt
from plotly import graph_objs as go
import plotly.express as px
import plotly.figure_factory as ff
from pylab import *
import matplotlib.patheffects as PathEffects

import descartes
import geopandas as gpd
from Levenshtein import distance
from itertools import product
from fuzzywuzzy import fuzz
from fuzzywuzzy import process
```

```
from scipy.spatial.distance import pdist, squareform
from shapely.geometry import Point, Polygon

import geoplot
from geopy.geocoders import Nominatim

import warnings
warnings.filterwarnings('ignore')
```

Data Exploration Highlights

```
# read & load the dataset into pandas dataframe
df = pd.read_csv('../input/us-accidents/US_Accidents_Dec21_updated.csv')
df.head(10)

#find missing values- How many?
list(df) #list column names
list(df.City)
# convert the Start_Time & End_Time Variable into Datetime Feature
df.Start_Time = pd.to_datetime(df.Start_Time)
df.End_Time = pd.to_datetime(df.End_Time)

# create a dataframe of city and their corresponding accident cases
city_df = pd.DataFrame(df['City'].value_counts()).reset_index().rename(columns={'ind

top_10_cities = pd.DataFrame(city_df.head(10)) #creates variable, called top_10_citi

fig, ax = plt.subplots(figsize = (12,7), dpi = 80)

cmap = cm.get_cmap('rainbow', 10)
clrs = [matplotlib.colors.rgb2hex(cmap(i)) for i in range(cmap.N)]

ax=sns.barplot(y=top_10_cities['Cases'], x=top_10_cities['City'], palette='rainbow')

total = sum(city_df['Cases'])
for i in ax.patches:
    ax.text(i.get_x()+.03, i.get_height()-4000, \
            str(round((i.get_height()/total)*100, 2))+'%', fontsize=15, weight='bold
                color='white')

plt.title('\nTop 10 Cities in US with most no. of \nRoad Accident Cases (2016-2021)\

plt.rcParams['font.family'] = "Microsoft JhengHei UI Light"
plt.rcParams['font.serif'] = ["Microsoft JhengHei UI Light"]

plt.ylim(1000, 120000)
plt.xticks(rotation=10, fontsize=12)
plt.yticks(fontsize=12)
```

```
ax.set_xlabel('\nCities\n', fontsize=15, color='grey')
ax.set_ylabel('\nAccident Cases\n', fontsize=15, color='grey')

for i in ['bottom', 'left']:
    ax.spines[i].set_color('white')
    ax.spines[i].set_linewidth(1.5)

right_side = ax.spines["right"]
right_side.set_visible(False)
top_side = ax.spines["top"]
top_side.set_visible(False)

ax.set_axisbelow(True)
ax.grid(color='#b2d6c7', linewidth=1, axis='y', alpha=.3)
MA = mpatches.Patch(color=clrs[0], label='City with Maximum\n no. of Road Accidents'
ax.legend(handles=[MA], prop={'size': 10.5}, loc='best', borderpad=1,
          labelcolor=clrs[0], edgecolor='white');
plt.show()

hightest_cases = city_df.Cases[0]
print(round(hightest_cases/6))
print(round(hightest_cases/(6*365)))
#Month with the most recorded accident

# US States
states = gpd.read_file('../input/us-states-map')

def lat(city):
    address=city
    geolocator = Nominatim(user_agent="Your_Name")
    location = geolocator.geocode(address)
    return (location.latitude)

def lng(city):
    address=city
    geolocator = Nominatim(user_agent="Your_Name")
    location = geolocator.geocode(address)
    return (location.longitude)

# list of top 10 cities
top_ten_city_list = list(city_df.City.head(10))

top_ten_city_lat_dict = {}
top_ten_city_lng_dict = {}
for i in top_ten_city_list:
    top_ten_city_lat_dict[i] = lat(i)
    top_ten_city_lng_dict[i] = lng(i)

top_10_cities_df = df[df['City'].isin(list(top_10_cities.City))]
```

```
top_10_cities_df['New_Start_Lat'] = top_10_cities_df['City'].map(top_ten_city_lat_di
top_10_cities_df['New_Start_Lng'] = top_10_cities_df['City'].map(top_ten_city_lng_di

geometry_cities = [Point(xy) for xy in zip(top_10_cities_df['New_Start_Lng'], top_10
geo_df_cities = gpd.GeoDataFrame(top_10_cities_df, geometry=geometry_cities)

fig, ax = plt.subplots(figsize=(15,15))
ax.set_xlim([-125,-65])
ax.set_ylim([22,55])
states.boundary.plot(ax=ax, color='grey');

colors = ['#e6194B','#f58231','#ffe119','#bfef45','#3cb44b', '#aaffc3','#42d4f4','#4
markersizes = [50+(i*20) for i in range(10)][::-1]
for i in range(10):
    geo_df_cities[geo_df_cities['City'] == top_ten_city_list[i]].plot(ax=ax, markers
                                                                      color=colors[i
                                                                      label=top_ten_

plt.legend(prop={'size': 13}, loc='best', bbox_to_anchor=(0.5, 0., 0.5, 0.5), edgeco

for i in ['bottom', 'top', 'left', 'right']:
    side = ax.spines[i]
    side.set_visible(False)

plt.tick_params(top=False, bottom=False, left=False, right=False,
                labelleft=False, labelbottom=False)

plt.title('\nVisualization of Top 10 Accident Prone Cities in US (2016-2021)', size=

fig, ax = plt.subplots(figsize=(15,15))
ax.set_xlim([-125,-65])
ax.set_ylim([22,55])
states.boundary.plot(ax=ax, color='grey');

colors = ['#e6194B','#f58231','#ffe119','#bfef45','#3cb44b', '#aaffc3','#42d4f4','#4
markersizes = [50+(i*20) for i in range(10)][::-1]
for i in range(10):
    geo_df_cities[geo_df_cities['City'] == top_ten_city_list[i]].plot(ax=ax, markers
                                                                      color=colors[i
                                                                      label=top_ten_

plt.legend(prop={'size': 13}, loc='best', bbox_to_anchor=(0.5, 0., 0.5, 0.5), edgeco

for i in ['bottom', 'top', 'left', 'right']:
    side = ax.spines[i]
    side.set_visible(False)

plt.tick_params(top=False, bottom=False, left=False, right=False,
                labelleft=False, labelbottom=False)
```

```
plt.title('\nVisualization of Top 10 Accident Prone Cities in US (2016-2021)', size=


ax.set(ylim =(-10000, 800000))
ax1.set(ylim =(-100000, 1700000))

plt.title('\nTop 10 States with most no. of \nAccident cases in US (2016-2021)\n', s
ax1.axes.yaxis.set_visible(False)
ax.set_xlabel('\nStates\n', fontsize=15, color='grey')
ax.set_ylabel('\nAccident Cases\n', fontsize=15, color='grey')

for i in ['top','right']:
    side1 = ax.spines[i]
    side1.set_visible(False)
    side2 = ax1.spines[i]
    side2.set_visible(False)

ax.set_axisbelow(True)
ax.grid(color='#b2d6c7', linewidth=1, axis='y', alpha=.3)

ax.spines['bottom'].set_bounds(0.005, 9)
ax.spines['left'].set_bounds(0, 800000)
ax1.spines['bottom'].set_bounds(0.005, 9)
ax1.spines['left'].set_bounds(0, 800000)
ax.tick_params(axis='y', which='major', labelsize=10.6)
ax.tick_params(axis='x', which='major', labelsize=10.6, rotation=10)

MA = mpatches.Patch(color=clrs[0], label='State with Maximum\n no. of Road Accidents
ax.legend(handles=[MA], prop={'size': 10.5}, loc='best', borderpad=1,
          labelcolor=clrs[0], edgecolor='white');
```

Initial Machine Learning Results - Partial Codes

```
# read & load the dataset into pandas dataframe
df = pd.read_csv('../input/us-accidents/US_Accidents_Dec21_updated.csv')

# Import KNeighborsClassifier from sklearn.neighbors
from sklearn.neighbors import KNeighborsClassifier

# Import DecisionTreeClassifier from sklearn.tree
from sklearn.tree import DecisionTreeClassifier

# Import RandomForestClassifier
from sklearn.ensemble import RandomForestClassifier

# Import LogisticRegression
from sklearn.linear_model import LogisticRegression

from sklearn.model_selection import train_test_split
```

```
from sklearn.model_selection import GridSearchCV
from sklearn.feature_selection import SelectFromModel
from sklearn.metrics import classification_report
from sklearn.metrics import confusion_matrix
from sklearn.metrics import accuracy_score
from sklearn.metrics import roc_curve, auc

#get some info
df.info()
# Convert Start_Time and End_Time to datetypes
df['Start_Time'] = pd.to_datetime(df['Start_Time'], errors='coerce')
df['End_Time'] = pd.to_datetime(df['End_Time'], errors='coerce')

# Extract year, month, day, hour and weekday
df['Year']=df['Start_Time'].dt.year
df['Month']=df['Start_Time'].dt.strftime('%b')
df['Day']=df['Start_Time'].dt.day
df['Hour']=df['Start_Time'].dt.hour
df['Weekday']=df['Start_Time'].dt.strftime('%a')

# Extract the amount of time in the unit of minutes for each accident, round to the
td='Time_Duration(min)'
df[td]=round((df['End_Time']-df['Start_Time'])/np.timedelta64(1,'m'))
df.info()

# Check if there is any negative time_duration values
df[td][df[td]<=0]

# Drop the rows with td<0

neg_outliers=df[td]<=0

# Set outliers to NAN
df[neg_outliers] = np.nan

# Drop rows with negative td
df.dropna(subset=[td],axis=0,inplace=True)
df.info()

# Remove outliers for Time_Duration(min): n * standard_deviation (n=3), backfill wit

n=3

median = df[td].median()
std = df[td].std()
outliers = (df[td] - median).abs() > std*n

# Set outliers to NAN
df[outliers] = np.nan
```

```
# Fill NAN with median
df[td].fillna(median, inplace=True)
df.info()

# Set the list of features to include in Machine Learning
feature_lst=['Severity','Start_Lng','Start_Lat','Distance(mi)','Side','City','County

# Set state
#state='NC'

# Select the state of North Carolina
df_state=df_sel.loc[df_sel.State=='NC'].copy()
df_state.drop('State',axis=1, inplace=True)
df_state.info()

# Map of accidents, color code by county

sns.scatterplot(x='Start_Lng', y='Start_Lat', data=df_state, hue='County', legend=Fa
plt.show()

# Generate dummies for categorical data
df_state_dummy = pd.get_dummies(df_state,drop_first=True)

df_state_dummy.info()

# Assign the data
df=df_state_dummy


# Set the target for the prediction
target='Severity'



# Create arrays for the features and the response variable

# set X and y
y = df[target]
X = df.drop(target, axis=1)

# Split the data set into training and testing data sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_stat

# List of classification algorithms
algo_lst=['Logistic Regression',' K-Nearest Neighbors','Decision Trees','Random Fore

# Initialize an empty list for the accuracy for each algorithm
accuracy_lst=[]

# Logistic regression
```

```
lr = LogisticRegression(random_state=0)
lr.fit(X_train,y_train)
y_pred=lr.predict(X_test)

# Get the accuracy score
acc=accuracy_score(y_test, y_pred)

# Append to the accuracy list
accuracy_lst.append(acc)

print("[Logistic regression algorithm] accuracy_score: {:.3f}.".format(acc))
```