12-19-2021

Analysis of Car Accidents Causes in the USA

Mohamed Aljaban
ma3982@rit.edu

RIT

Analysis of Car Accidents Causes in the USA

By

Mohamed Aljaban

A Capstone Submitted in Partial Fulfillment of the Requirements for the Degree of Master of Science in Professional Studies:

Data Analytics

Department of Graduate Programs & Research

Rochester Institute of Technology

RIT Dubai

19 December 2021

RIT

Master of Science in Professional Studies:

Data Analytics

Graduate Capstone Approval

Student Name: Mohamad Aljaban

Graduate Capstone Title: Analysis of Car Accidents Causes in the USA

Graduate Capstone Committee:

Name: Dr. Sanjay Modak **Date:**

Chair of committee

Name: Dr. Ehsan Warriach **Date:**

Member of committee

Acknowledgments

I want to start this by thanking Rochester Institute of Technology for providing this valuable and interesting program to the students. I would also like to thank all the instructors who taught me and delivered to me valuable information. A sincere thanks to Dr. Mick McQuaid who taught me the essential of R the programming language, and my mentor Dr. Ehsan Warriach. And I would like to greet Dr. Sanjay Modak, our chair of committee, for running this program, and thank him for giving us all the support we needed, and providing us with all the material needed to accomplish this mission. And finally to my best friend Majd Ibrahim for motivating me to enroll in this program and supporting me throughout this incredible journey of learning and growth.

Abstract

Over the past decade more people are opting in to purchase and own their own personal vehicle and in some instances several, which means that road accidents rates are doomed to increase. And this presents a challenge to the government, individuals and the collective community as car accidents are in most cases life threatening and a hazard to society. Thus, this paper aims to tackle this issue and dig deep to explore the main factors contributing to the increase of car accidents rate. The dataset used in this research is data collected from traffic accidents events captured by the department of transportation, law-enforcement agencies, and traffic cameras continuously in the United States from 2016 to 2020. Two models were performed to predict the impact of car accidents on road traffic, with a focus on the leading factors contributing to road accidents. Results showed that the main two factors affecting car accidents rate are traffic caused by work rush hour and population density. Furthermore, this research can be used to create solutions to limit and decrease car accidents in cities, such as adopting the working from home concept, facilitating the ownership of self-driving vehicles, creating a seamless public transportation infrastructure, and distributing rush hours throughout the day to name a few.

Keywords: Road Accidents, Traffic, Traffic Severity, Human errors, Rush-hour

Table of Contents

List of Tables:

List of Figures

Chapter 1

1.1 Background Information

Cars adoption has been increasing dramatically in recent years due to the rapid technology development and the fast paced life individuals in cities live. According to the U.S department of transportation (DOT, 2021), the number of driving licenses issued in 2020 was 228,687 compared to 190,625 driving licenses issued in 2010, which means there are more individuals and cars on the road now than a decade ago, thus leading to a higher car accident rate as the relationship is linear.

Car accidents can take many forms and they occur due to so many reasons; some are controllable, such as speed, tailgating, wrong lane changing, sleep, or reckless driving; and others are not, such as car failure, bad weather conditions, unavailability of road signs, or bad road conditions. But the most frequent form of car accidents is car collisions which can be a single-car accident, rear-end collisions, sideswipe collisions, multiple vehicle collisions, or car rollovers. This can have serious negative consequences such as injuries, disabilities, and even in many cases death. Not to mention property loss and damage of the vehicle, giving rise to both personal and social losses. According to the world health organization (WHO, 2020), 1.3 million deaths and around 50 million injuries happen every year due to crashes. The french epidemiological center (Ce´ piDC, 2015) stated that individuals aged 15-24 had the highest average mortality rate. Governments, insurance companies, employers, and individuals suffer financially because of accidents; the national highway traffic safety administration (NHTSA, 2010) stated that estimated losses due to accidents could reach $1 trillion due to loss of productivity and loss of life.

1.2 Statement of the problem

Car accidents impose a serious threat on the livelihood and financial hood of citizens and the government; millions of people worldwide have lost their lives due to accidents or suffered severe injuries and disabilities, affecting their own quality of lives and their loved ones.

These accidents negatively affect the country's productivity, its GDP and tax receivables. Governments, employers, and insurance companies bear significant financial losses by covering medical expenses and compensating those affected by these accidents. However, causes of accidents vary, and contributors differ for several reasons; previous research papers did not include a large sample size and excluded certain conditions, resulting in incomplete analysis and conclusion.

1.3 Project goals

The goal of this paper is to provide the necessary information for government officials to build valuable solutions to address the rapid increase of car accidents and limiting it to a minimum. Thus, creating a safer environment for everyone by addressing the causes of these life threatening accidents and in return reducing the financial burden imposed on the government, individuals and businesses.

The main objective of the project is to analyze all the available factors and gain an in-depth understanding of the factors contributing to car accidents.

The questions of this research paper are the following:

1. Are road accidents increasing, and what is the rate of this increase?
2. Where and when do most road accidents take place?
3. What are the possible solutions to address road accidents?

1.4 Methodology

There are many factors contributing to road accidents, and to be able to pinpoint the main reasons causing those accidents, a thorough study needs to be conducted to build a comprehensive understanding of how each factor affects the rate in which road accidents increase.

CRISP-DM presents an excellent methodology to follow when doing a data analytics project. The first step of this methodology is **Data understanding**: the data is a collection of accidents that happened in the United States from the period 2016 to 2020 and consist of 3 million observations and 47 different attributes. The data is well structured, organized and cleaned. The second step is **Data Preparation**, the data here is being modified through removing unwanted attributes, and adding new attributes that will be helpful during the evaluation phase, in addition to dealing with missing values. Once preparation is done, different types of visualizations are performed to understand the relationship between the variables and accidents. The third step is the **Modeling**, after performing the visualizations and understanding the relevance between accidents and the chosen attribute, two different algorithms will be performed and compared to predict the traffic severity, which are the Random Forest, and the Naive Bayes. The last step will consist of **Evaluating** the models and visualization results. Once the results are concluded, concerned parties can use these results to build effective solutions to address the main problem which is the rapid increase in road accidents.

1.5 Limitations of the study

The limitation imposed on this research paper stems from the incomplete information provided by the dataset; the dataset does not contain any information about the profile of the driver, such as their age, gender, race, or their state of consciousness. These details can help us a lot when identifying the factors for road accidents and can help us detect a pattern.

Moreover, the dataset lacked vehicle information, such as, car specifications (sedan, sports car, truck, bus), or the type of ownership (private, public or rented). All these additional pieces of information could have given us better insights of the data set.

Chapter 2 - Literature review

Billions of dollars are spent every year to cover car crashes. A study was conducted by (Miller, Bhattacharya, Zaloshnja, and Taylor, 2011); governments pay approximately 35 billion dollars every year covering both medical expenses and social welfare to injured individuals, in addition to the forgone taxes resulting from the injuries or death of individuals.

Road traffic crashes comprise leading economic and health challenges, especially for developing countries (Chen, 2010). Factors causing accidents are many; according to (Reyner and Horne, 1998), a large volume of road accidents happen because of drivers falling asleep, and many of those accidents are related to work. A study conducted by (Knipling and wang, 1994) stated that every lorry would be at least involved in one related sleep crash in the United States. (Reyner and Horne, 1998) Found that male drivers below 30 are susceptible to sleep-related accidents in the early morning due to destructive sleeping patterns; however, later in the afternoon, sleep-related accidents shift to drivers aged 50 years and above. (Reyner and Horne, 1998) concluded that self-awareness is the key to prevent such incidents, and sleeping detectors will not help; the best solution is to pull over and stop driving.

(Celik and Oktay, 2014) argued that less educated drivers are more vulnerable to fatal accidents. Their study showed that the driver's age and time of driving are strong predictors of fatal accidents. According to (Tadege, 2020) the age of the drivers is the strong contributing factor behind fatal car accidents. Moreover, driver inexperience increases the proportion of human error leading to severe or fatal accidents. In addition, male drivers caused 254 out of 255 fatal accidents which is around 99.6% of accidents (Tadege, 2020). (Abu Jadayil, Khraisat, Shakoor,

2020) explains that young male drivers are more likely to cause an accident than older adults because of their tendencies to be more reckless, and exceed the speed limits which increases the chances of crashing.

Two main factors are associated with car accidents: traffic and human error (Gicquel, Ordonneau, Blot, Toillon, Ingrand, and Romo, 2017). Human errors can take many forms, such as wrong decision-making, sleeping, tailgating, alcohol and drugs consumption, mobile usage, and more. According to (Abu Jadayil, Khraisat, Shakoor, 2020), human error is the primary cause of accidents where 96.8% out of 97,981 accidents happen because of it. (Fan, 2015) says that most accidents happen because of over-speeding, and he claimed that drivers could not avoid accidents, where time is minimal and decisions are hard to make. Therefore accidents occur. (Waylen and McKenna, 2008) stated that driving under the impact of drugs or alcohol triples the chances of having car accidents, decreasing the drivers' concentration, reflexes, and awareness level. The remarkable development of mobile phones has significantly changed the world, yet it has become a significant source of distraction. Texting and driving are becoming tremendously noticeable, especially for young people, causing many severe accidents. A study conducted by (Saifuzzaman, Haque, Zheng, and Washington, 2015) found that young drivers are highly distracted by mobile phones impacting their driving performance leading to car crashes. Drivers who use their phones do not pay attention to the car in front of them, especially when the car in front slows down, or even sometimes drivers' might not pay attention to road signs, such as traffic signals or stop signs. Accidents can also happen due to immoderate weather conditions, road conditions, or a car breakdown. (Fan, 2015) concluded that brakes failure, steering system failure, car light failure, or tire burst are unavoided factors that lead to car accidents; however,

the likelihood of this happening is very rare. (Chen, Zhao, Liu, Ren, and Liu, 2019) argued that weather and road conditions have a significant impact on driver's behavior, putting the drivers in a critical position that can cause accidents. Snow and dense fog tend to affect drivers' behavior due to low visibility. In addition (Mao, Yuan, Gan, and Zhang, 2019) emphasized that driver age and gender, weather condition, traffic density, vehicle speed, lane change behavior, vehicle type, time of the day, and day of the week are all important factors affecting car accidents. Road lighting and visibility are other vital factors that affect the driver's behavior. (Farooq and Juhasz, 2019) stated that driver's visibility is an essential factor when it comes to car accidents. The two main contributing factors that affect driver visibility and reaction resulting in a car accident are mobile phones, and blind spots, in addition to the fact that drivers cannot brake or avoid the collision when visibility issues are observed. According to (Boyce 2003), darkness reduces visibility and is associated with a higher degree of perceptual errors, including distraction and lack of attention. A study performed by (Jägerbrand and Sjöbergh, 2016) claimed that vehicle speed in clear weather conditions and daylight is higher than in the hours of darkness. Moreover, rain significantly affected fatalities and serious injuries; however, rush hours and extreme night conditions were excluded from the study.

The reality is car accidents are a serious issue that affects people’s lives, aside from the tremendous amount of money paid by governments and institutions, covering the damages of those accidents. And the factors that lead to them, such as sleeping while driving, using mobile phones, poor road and lighting conditions, or bad weather conditions are apparent. Moreover, as the previous research showed it was proven that age, gender, and human errors are the most influencing factors that might contribute to a road accident. However, these research papers are

not quite complete because of the limitations they had such as the small sample size, and missing information about traffic hours, vehicle speed, location, extreme weather conditions, and drivers profile.

Chapter 3 - Project Description

3.1 Data Collection

Data source, quality, and reliability are the most important factors when it comes to data analysis. The data set has been collected from several providers using an application programming interface (APIs) that broadcasts traffic accidents events captured by the department of transportation, law-enforcement agencies, and traffic cameras continuously since February 2016, covering 49 states in the United States. It has 3 million observations and 47 attributes.

3.2 Data Exploration

The dataset contains 47 different attributes; table 1 includes the name of the attributes, description, and type. *Figure 1,2*, and *3* represents a data summary that illustrates all the attributes along with their components.

#	Attribute	Description	Type
1	ID	A number that identify the accident record	Discrete
2	Severity	A number that represents the severity of the accidents on traffic a scale of 1 represents a low impact, and 4 represents severe impact	Nominal
3	Start time	A date that represents start time of the accident	Continuous
4	End time	A date that represents end time of the accident	Continuous
5	Starting latitude	Represents latitude in GPS coordinates of the starting point	Continuous
6	Starting longitude	Represents longitude in GPS coordinates of the starting point	Continuous
7	Ending latitude	Represents latitude in GPS coordinates of the ending point	Continuous
8	Ending longitude	Represents longitude in GPS coordinates of the ending point	Continuous
9	Distance	The length of the road affected by the accident	Continuous

10	Description	A description of the accident written in natural language	Nominal
11	Number	Represents the street number	Continuous
12	Street	Represents the street name	Nominal
13	Side	Describe where the accident took place, Right or Left lane	Nominal
14	City	Represent the city of which the accident took place	Nominal
15	County	Represent the county of which the accident took place	Nominal
16	State	Represent the state of which the accident took place	Nominal
17	Zip Code	Represents the Zipcode of the address	Nominal
18	Country	Represent the country of which the accident took place	Nominal
19	Timezone	Shows the timezone of which the accident took place	Nominal
20	Airport code	Shows the code of the airport closest to the accident location	Nominal
21	Weather timestamp	Represents the date and time the accident took place	Continuous
22	Temperature	A number that represents the temperature in fahrenheit	Continuous
23	Wind Chill	A number that represents the wind chill in fahrenheit	Continuous
24	Humidity	A number that represents humidity in (%)	Continuous
25	Pressure	A number that represents air pressure in inches	Continuous
26	Visibility	A number that shows visibility in miles	Continuous
27	Wind direction	Represents the wind direction	Nominal
28	Wind speed	A number that represents the wind speed in miles per hour	Continuous
29	precipitation	Represents precipitation in inches	Continuous
30	Weather condition	States the status of the weather (clear, rainy, snowy, etc)	Nominal
31	Amenity	A point of interest annotation which indicates the presences of Amenity in a nearby location, addressed as True or False	Nominal
32	Bump	A point of interest annotation which indicates the presences of Bump in a nearby location, addressed as True or False	Nominal
33	Crossing	A point of interest annotation which indicates the presences of crossing in a nearby location, addressed as True or False	Nominal
34	Give way	A point of interest annotation which indicates the presences of give way in a nearby location, addressed as True or False	Nominal
35	Junction	A point of interest annotation which indicates the presences of junction in a nearby location, addressed as True or False	Nominal
36	No exit	A point of interest annotation which indicates the presences of no exit in a nearby location, addressed as True or False	Nominal

37	Railway	A point of interest annotation which indicates the presences of railway in a nearby location, addressed as True or False	Nominal
38	Roundabout	A point of interest annotation which indicates the presences of roundabout in a nearby location, addressed as True or False	Nominal
39	Station	A point of interest annotation which indicates the presences of station in a nearby location, addressed as True or False	Nominal
40	Stop	A point of interest annotation which indicates the presences of stop in a nearby location, addressed as True or False	Nominal
41	Traffic calming	A point of interest annotation which indicates the presences of traffic calming in a nearby location, addressed as True or False	Nominal
42	Traffic signal	A point of interest annotation which indicates the presences of traffic signal in a nearby location, addressed as True or False	Nominal
43	Turning loop	A point of interest annotation which indicates the presences of turning loop in a nearby location, addressed as True or False	Nominal
44	Sunrise-Sunset	represents the period of the day (day or night) based on sunrise and sunset	Nominal
45	Civil twilight	represents the period of the day (day or night) based on civil twilight	Nominal
46	Nautical twilight	represents the period of the day (day or night) based on Nautical twilight	Nominal
47	Astronomical twilight	represents the period of the day (day or night) based on Astronomical Twilight	Nominal

Table 1: List of attributes and their description

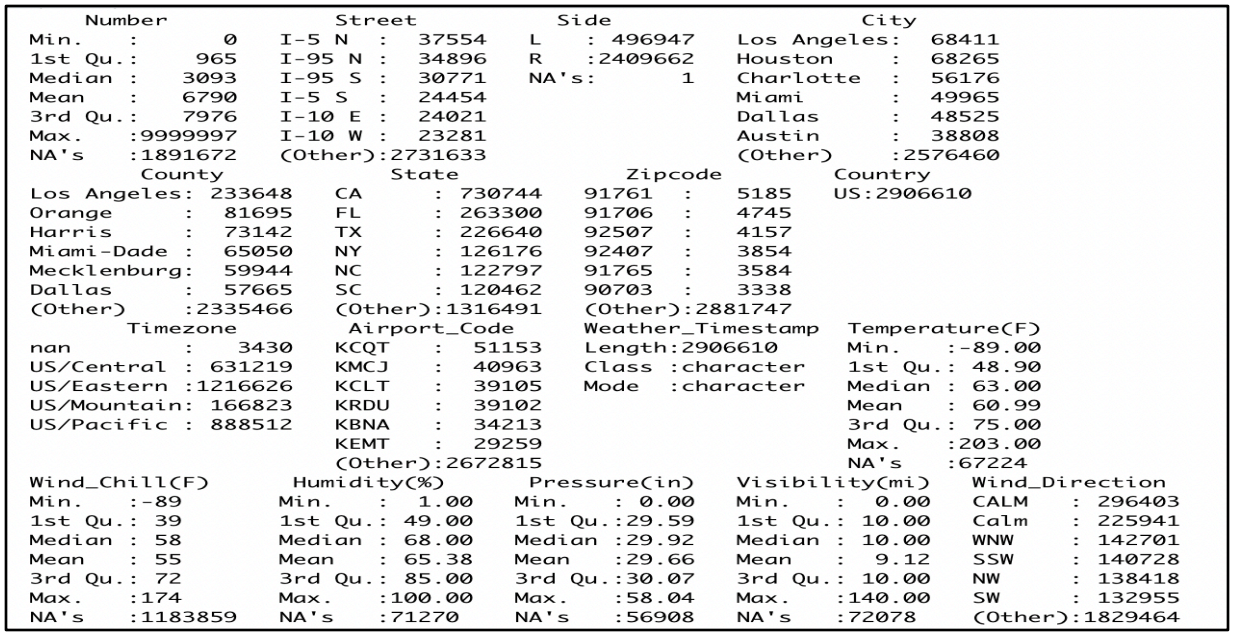
Chapter 4 Project Analysis

4.1 Exploratory Data Analysis

Figure 1, 2, and *3* represents the summary of the attributes. Time, Latitude, Longitude, distance, number, temperature, humidity, pressure, visibility, wind direction, wind speed, and precipitation are all numeric variables, therefore R list them from least to greatest and display their five number summary minimum: the lowest number in the dataset, first quartile represents the 25% value, median represents the 50% value, third quartile represents the 75% value, and Maximum, represents the Highest numbers in the dataset, in addition to the mean which represents the average. Severity, description, street, side, city, county, state, zip code, country, timezone, airport code, weather conditions, sunrise-sunset, nautical twilight, civil twilight, and astronomical twilight are all structured as a factor, therefore R displays the values available in each variable with the count of how many times the value occurs in the dataset . Amenity, bump, crossing, give-away, junction, no exit, roundabout, traffic signal, turning loop are logical expressions, R display how many True or False outcomes are there for each variable.

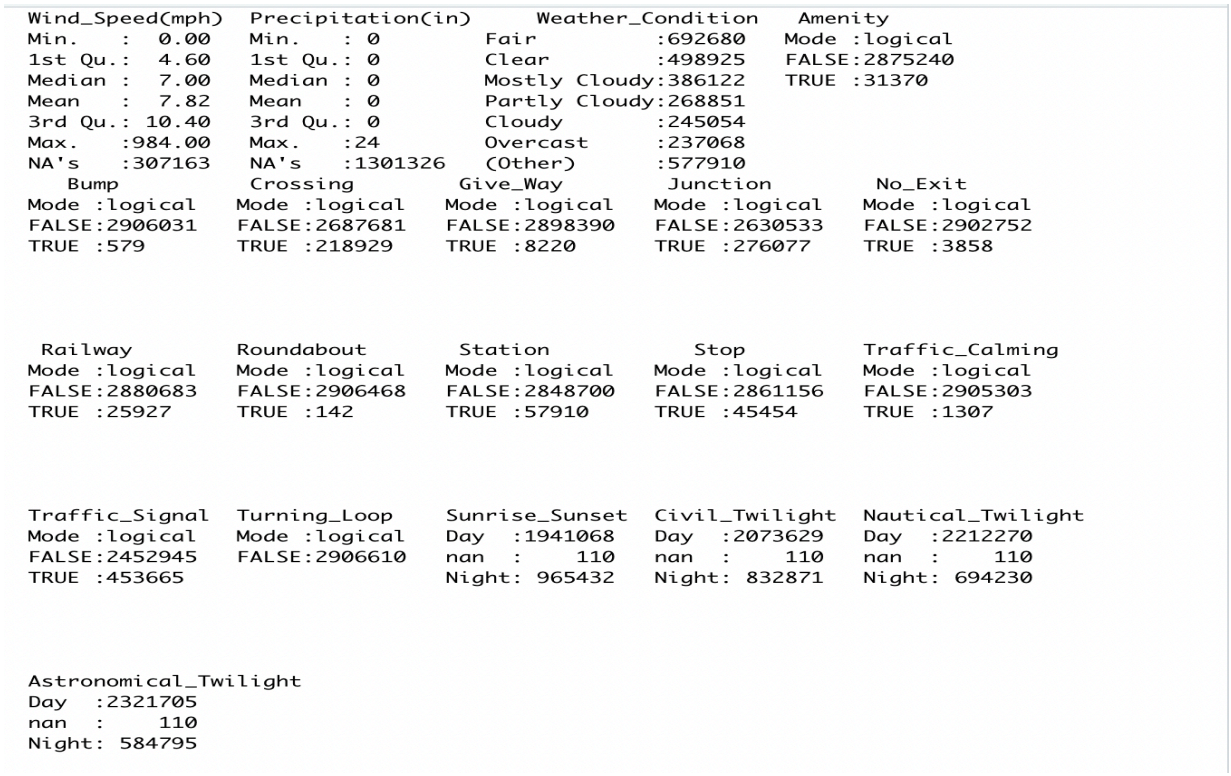
```
      ID              Severity       Start_Time
Length:2906610      1:   28751   Min.   :2016-02-08 00:37:08
Class :character    2:2129263   1st Qu.:2018-01-05 14:59:59
Mode  :character    3: 629452   Median :2019-05-12 00:43:00
                    4: 119144   Mean   :2019-02-25 19:48:23
                                3rd Qu.:2020-05-15 14:04:51
                                Max.   :2020-12-31 23:28:56

   End_Time                       Start_Lat       Start_Lng          End_Lat
Min.   :2016-02-08 06:37:08   Min.   :24.56   Min.   :-124.62   Min.   :24.56
1st Qu.:2018-01-05 17:26:06   1st Qu.:33.66   1st Qu.:-117.82   1st Qu.:33.65
Median :2019-05-12 03:28:00   Median :36.10   Median : -91.17   Median :36.06
Mean   :2019-02-25 22:40:12   Mean   :36.53   Mean   : -96.43   Mean   :36.52
3rd Qu.:2020-05-15 15:20:41   3rd Qu.:40.38   3rd Qu.: -80.86   3rd Qu.:40.33
Max.   :2021-01-01 00:00:00   Max.   :49.00   Max.   : -67.11   Max.   :49.08
                                                                NA's   :282821
   End_Lng          Distance(mi)
Min.   :-124.62   Min.   :  0.0000
1st Qu.:-117.70   1st Qu.:  0.0000
Median : -91.05   Median :  0.0000
Mean   : -96.20   Mean   :  0.3981
3rd Qu.: -80.85   3rd Qu.:  0.2790
Max.   : -67.11   Max.   :333.6300
NA's   :282821
                                                          Description
A crash has occurred causing no to minimum delays. Use caution.:   2709
At I-15 - Accident.                                             :   2124
At I-5 - Accident.                                              :   1929
At I-405/San Diego Fwy - Accident.                              :   1782
At I-605 - Accident.                                            :   1492
At Grand Ave - Accident.                                        :   1117
(Other)                                                         :2895457
```

Figure 1: Data summary



```
     Number             Street          Side                  City
Min.   :       0   I-5 N  :  37554   L    : 496947   Los Angeles:  68411
1st Qu.:     965   I-95 N :  34896   R    :2409662   Houston    :  68265
Median :    3093   I-95 S :  30771   NA's:       1   Charlotte  :  56176
Mean   :    6790   I-5 S  :  24454                  Miami      :  49965
3rd Qu.:    7976   I-10 E :  24021                  Dallas     :  48525
Max.   :9999997   I-10 W :  23281                  Austin     :  38808
NA's   :1891672   (Other):2731633                  (Other)    :2576460
        County              State              Zipcode         Country
Los Angeles: 233648   CA      : 730744   91761  :   5185   US:2906610
Orange     :  81695   FL      : 263300   91706  :   4745
Harris     :  73142   TX      : 226640   92507  :   4157
Miami-Dade :  65050   NY      : 126176   92407  :   3854
Mecklenburg:  59944   NC      : 122797   91765  :   3584
Dallas     :  57665   SC      : 120462   90703  :   3338
(Other)    :2335466   (Other):1316491   (Other):2881747
       Timezone         Airport_Code     Weather_Timestamp  Temperature(F)
nan         :   3430   KCQT   :  51153   Length:2906610     Min.   :-89.00
US/Central : 631219   KMCJ   :  40963   Class :character   1st Qu.: 48.90
US/Eastern :1216626   KCLT   :  39105   Mode  :character   Median : 63.00
US/Mountain: 166823   KRDU   :  39102                      Mean   : 60.99
US/Pacific : 888512   KBNA   :  34213                      3rd Qu.: 75.00
                      KEMT   :  29259                      Max.   :203.00
                      (Other):2672815                      NA's   :67224
Wind_Chill(F)      Humidity(%)      Pressure(in)   Visibility(mi)   Wind_Direction
Min.   :-89       Min.   :  1.00   Min.   :  0.00   Min.   :  0.00   CALM   : 296403
1st Qu.: 39       1st Qu.: 49.00   1st Qu.:29.59   1st Qu.: 10.00   Calm   : 225941
Median : 58       Median : 68.00   Median :29.92   Median : 10.00   WNW    : 142701
Mean   : 55       Mean   : 65.38   Mean   :29.66   Mean   :  9.12   SSW    : 140728
3rd Qu.: 72       3rd Qu.: 85.00   3rd Qu.:30.07   3rd Qu.: 10.00   NW     : 138418
Max.   :174       Max.   :100.00   Max.   :58.04   Max.   :140.00   SW     : 132955
NA's   :1183859   NA's   :71270    NA's   :56908   NA's   :72078    (Other):1829464
```

Figure 2: Data summary



```
Wind_Speed(mph)  Precipitation(in)     Weather_Condition   Amenity
Min.   :  0.00   Min.   : 0         Fair          :692680   Mode :logical
1st Qu.:  4.60   1st Qu.: 0         Clear         :498925   FALSE:2875240
Median :  7.00   Median : 0         Mostly Cloudy:386122   TRUE :31370
Mean   :  7.82   Mean   : 0         Partly Cloudy:268851
3rd Qu.: 10.40   3rd Qu.: 0         Cloudy        :245054
Max.   :984.00   Max.   :24         Overcast      :237068
NA's   :307163   NA's   :1301326    (Other)       :577910
   Bump           Crossing         Give_Way         Junction         No_Exit
Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:2906031   FALSE:2687681   FALSE:2898390   FALSE:2630533   FALSE:2902752
TRUE :579       TRUE :218929    TRUE :8220      TRUE :276077    TRUE :3858


 Railway         Roundabout        Station           Stop          Traffic_Calming
Mode :logical   Mode :logical   Mode :logical   Mode :logical   Mode :logical
FALSE:2880683   FALSE:2906468   FALSE:2848700   FALSE:2861156   FALSE:2905303
TRUE :25927     TRUE :142       TRUE :57910     TRUE :45454     TRUE :1307


Traffic_Signal  Turning_Loop    Sunrise_Sunset  Civil_Twilight  Nautical_Twilight
Mode :logical   Mode :logical   Day  :1941068   Day  :2073629   Day  :2212270
FALSE:2452945   FALSE:2906610   nan  :    110   nan  :    110   nan  :    110
TRUE :453665                    Night: 965432   Night: 832871   Night: 694230


Astronomical_Twilight
Day  :2321705
nan  :    110
Night: 584795
```

Figure 3: Data Summary

4.2 Data preprocessing

The first step when doing analysis is the preprocessing stage which consists of multiple steps that will enhance the performance of the analysis. First is **data sampling**; as mentioned previously, the dataset is large; therefore, working with sample size is more convenient. A 20% is selected to work with for the analysis process.

Second is **data cleaning**; all datasets include missing values; therefore, dealing with them is necessary to enhance the performance of the analysis. *Figure 4* represents the missing values available in our data set; only three attributes: number, precipitation, and wind chill, have high NA's; therefore, they were removed.

Third is **data selection**; dropping ineffective attributes will ensure focusing on specific variables and extracting the maximum results; the ID variable represents the sequence, so it is not an important variable to keep. Longitude and latitude represent the geographic location. However, we have the city and state; therefore, they are not essential, zip code as well does not indicate anything relative, country all accident occurred in the United States, so it is ineffectual to include, airport code as well does not indicate anything relative, and weather timestamp is a duplicate of the start time of the accident; therefore all those variables were omitted.

Fourth is **data transformation**; all the nominal attributes severity, city, state, weather conditions, sunrise-sunset, side, and street were classified as character form, thus transferring them into factor form is necessary to have a deeper understanding of the insights for each attribute as well as the starting time of the accident was transformed to date format.

Fifth is **feature engineering** is one of the most important parts because several new attributes are created from existing ones, providing more insights about the data, enabling the idea of deriving

more decisive conclusions. The first column is Road type with an outcome of City or Highway, extracted from the street column. Moreover, Year, Month, day, and hour, was extracted from the start time column.

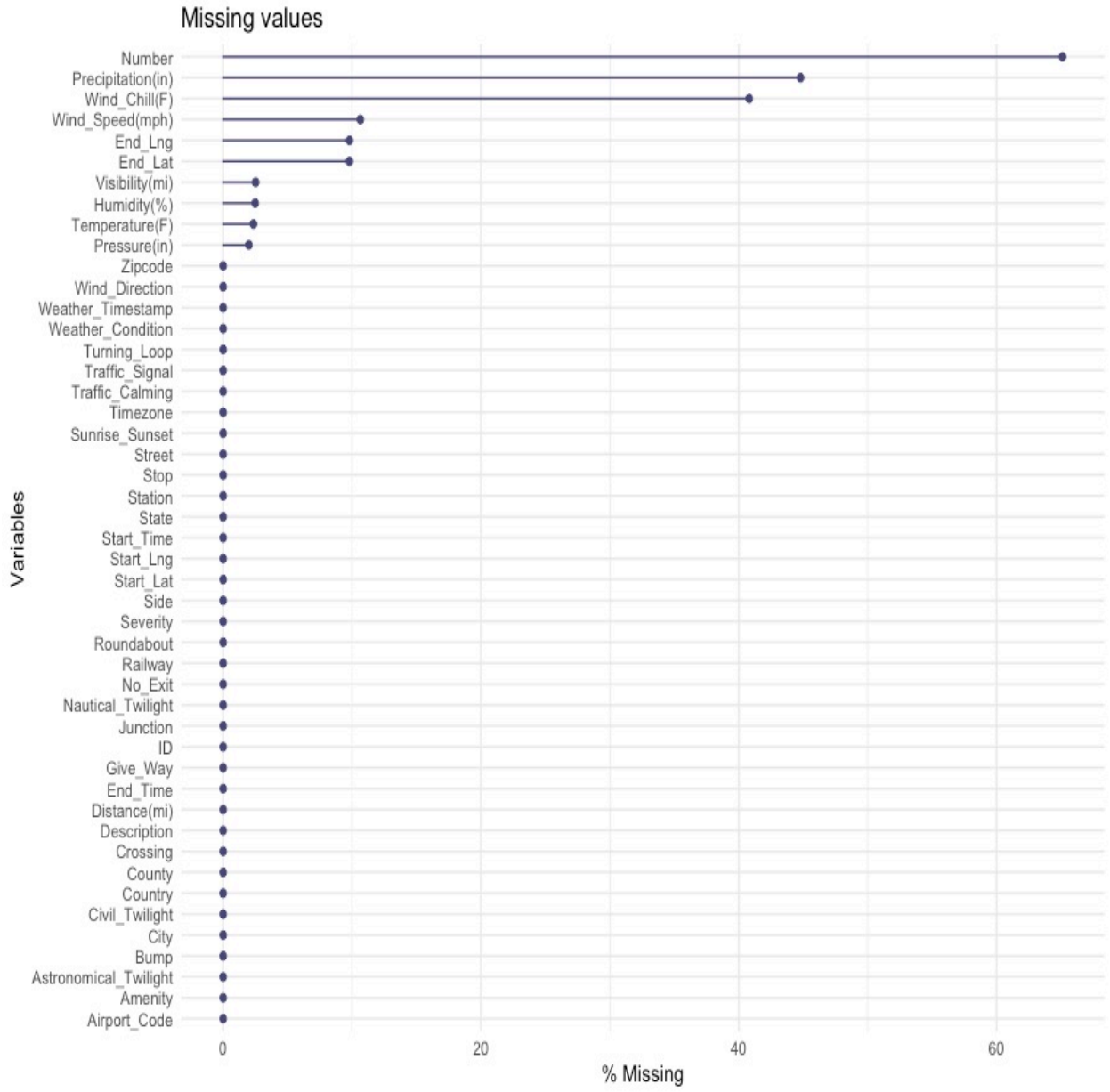


Figure 4: Missing values

4.3 Visualizations

4.3.1 Analysis of Time and Location

Feature engineering was performed on the start time column to extract the years, months, days, and hours.

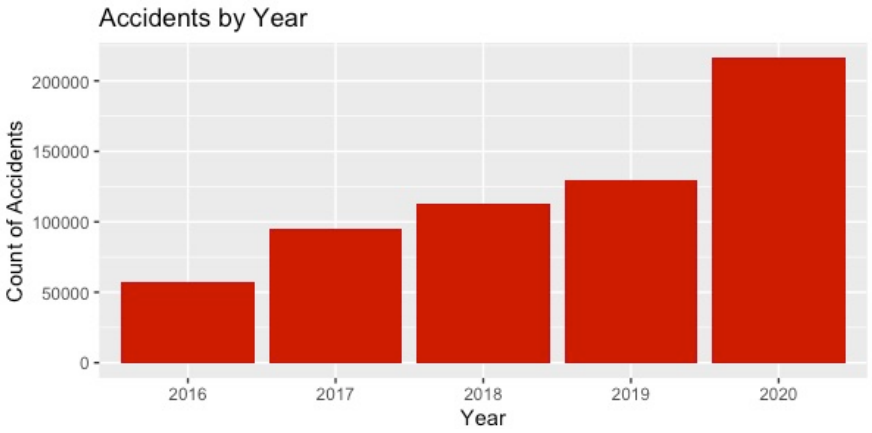


Figure 5: Distribution of accidents through Years

Figure 5 shows that car accidents have been increasing from year to year. This could be attributed to more people getting driving licenses, in addition to the fact that governments and relevant authorities are failing to tackle this issue and work on creating valuable solutions to decrease it.

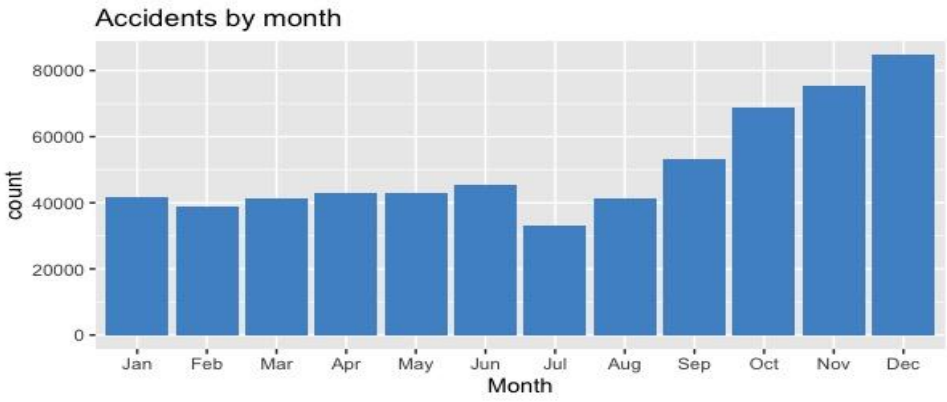


Figure 6: Distribution of accidents through Months

Figure 6 represents accidents distribution through months. The lowest rate occurs in July and August because it is the summer season, students are on vacation, and employees are on leave. Moreover, the highest ratio occurs at the end of the year which can be attributed to it being the end of year and the holiday season which is a busy season for business.

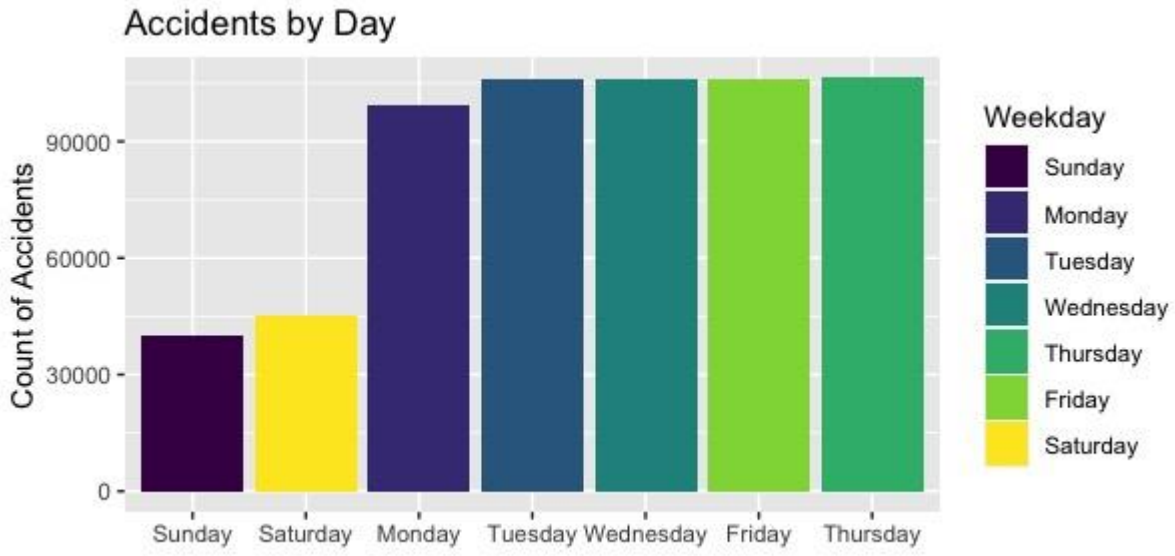


Figure 7: Distribution of accidents by Days

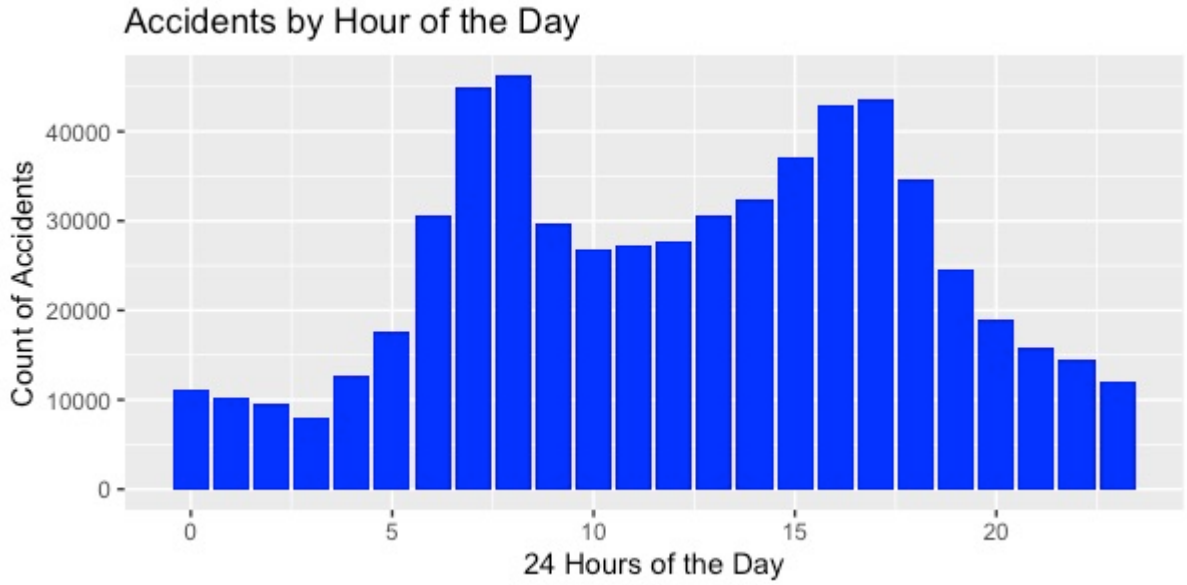


Figure 8: Distribution of accidents by Hours

Figure 7 and *Figure 8* are the most interesting, and they are highly correlated; looking at *Figure 4* shows that accidents are high on weekdays and low on weekends. *Figure 5* supports *Figure 4*, showing that the density of accidents occur the most at 7 and 8 am as well as 4 and 5 pm, which means that the majority of accidents occur at rush hour when people are either going to work or going back home. This leads us to the conclusion that traffic is one of the major causes of accidents because people drop their attention and lose their concentration in traffic by using mobile phones, and falling asleep while driving. Performing feature engineering revealed many insights that will help in interpretations and analysis. The next illustrations will give us a broader picture about where did accidents occur:

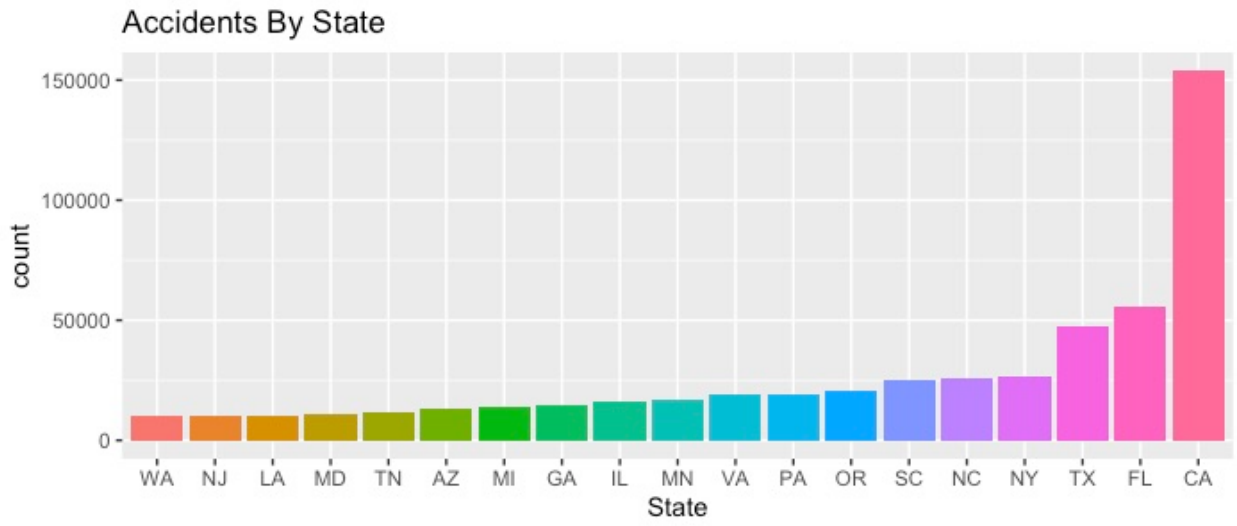


Figure 9: Distribution of accidents By State

Washingto	New Jersey	Louisiana	Maryland	Tennessee	Arizona	Michiga	Georgi
Illinois	Minnesota	Virginia	Pennsylvania	Oregon	South Carolina	North Carolina	
New York	Texas	Florida	California				

Table 2: Abbreviation Explanation

Figure 9 illustrates the distribution of accidents among states; it shows that California has the highest number of accidents, followed by Florida, and then Texas, then the rest has almost equal distribution of accidents. According to the (U.S census bureau, 2020), the population division showed that California has a population of almost 40M, Florida has almost 30M. Texas has 21M, which explains the distribution of accidents, indicating a positive correlation; the higher the population, the more the accidents will occur. The following Figure shows the city in which accidents took place. However, results showed that accidents occur in more than 9000 cities, which is hard to visualize; therefore, the number of accidents was filtered to 3500 accidents and above.

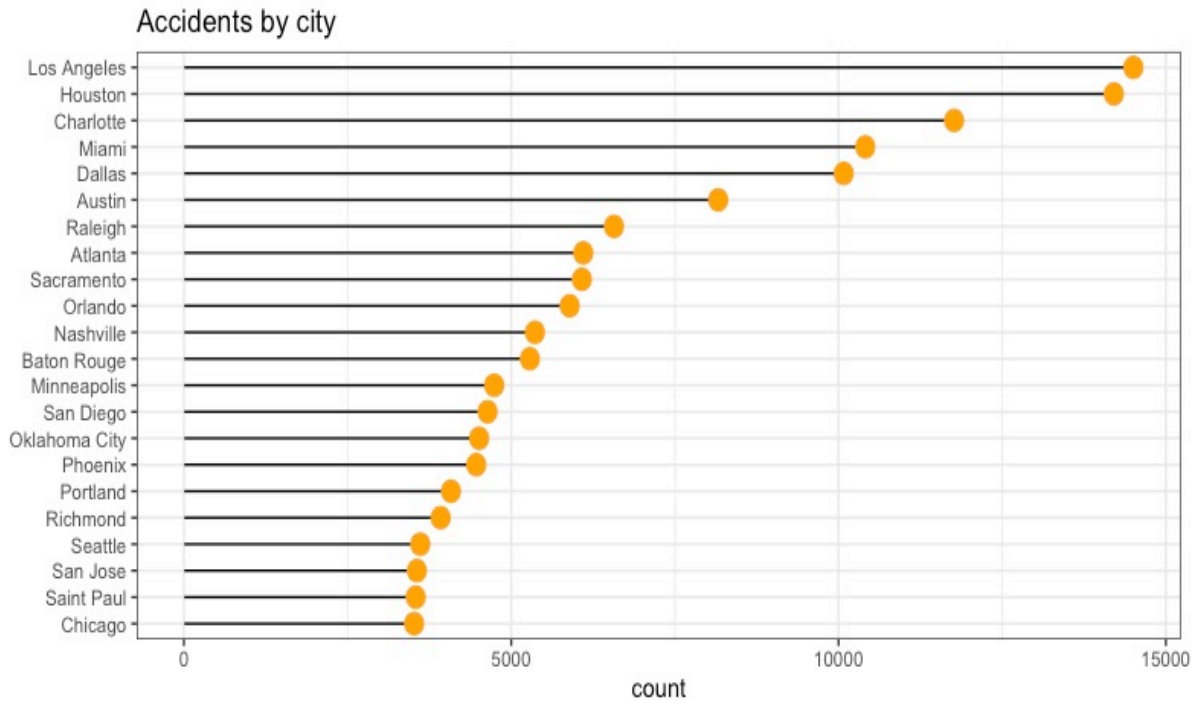


Figure 10: Distribution of accidents By City

Figure 10 is a subset of *Figure 9*, representing the cities; the highest is in Los Angeles, California, followed by Houston, TX then Charlotte, NC, then comes Miami, FL, and finally Dallas, TX.

4.3.2 Analysis of the Remaining Variables

After analyzing the new variables resulting from the feature engineering, the remaining variables were analyzed, illustrating figures related to the time of the day, weather condition, road type, and side. According to many researches conducted, authors argued that at night drivers are more vulnerable to exceeding the speed limit and driving recklessly. In addition to the negative impact that poor vision might impose on drivers. However, *Figure 11* shows that accidents occur in day time more, which harmonizes with the results of *Figure 5*.

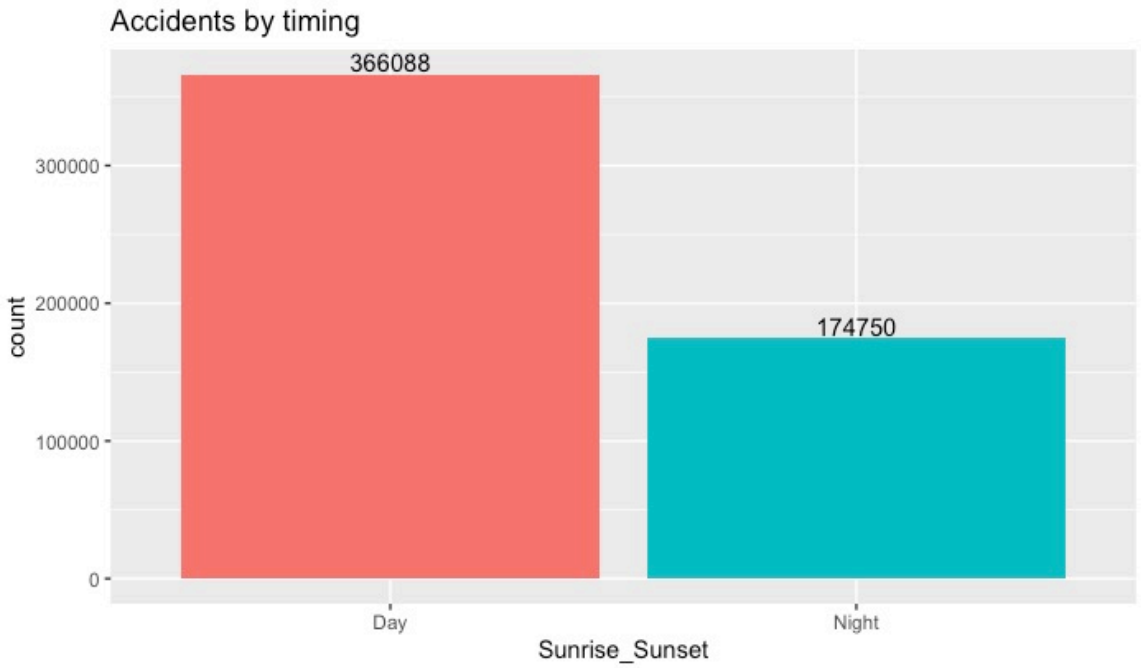


Figure 11: Distribution of Accidents By Day or Night

Many researchers have considered weather conditions, discussing the impact of rain, fog, wind, or snow on both road conditions and driver behavior. Nevertheless, results showed that weather conditions have no impact on road accidents. According to *Figure 12*, most accidents occur in clear, fair, or cloudy weather, denoting that weather conditions have a low impact on car accidents.

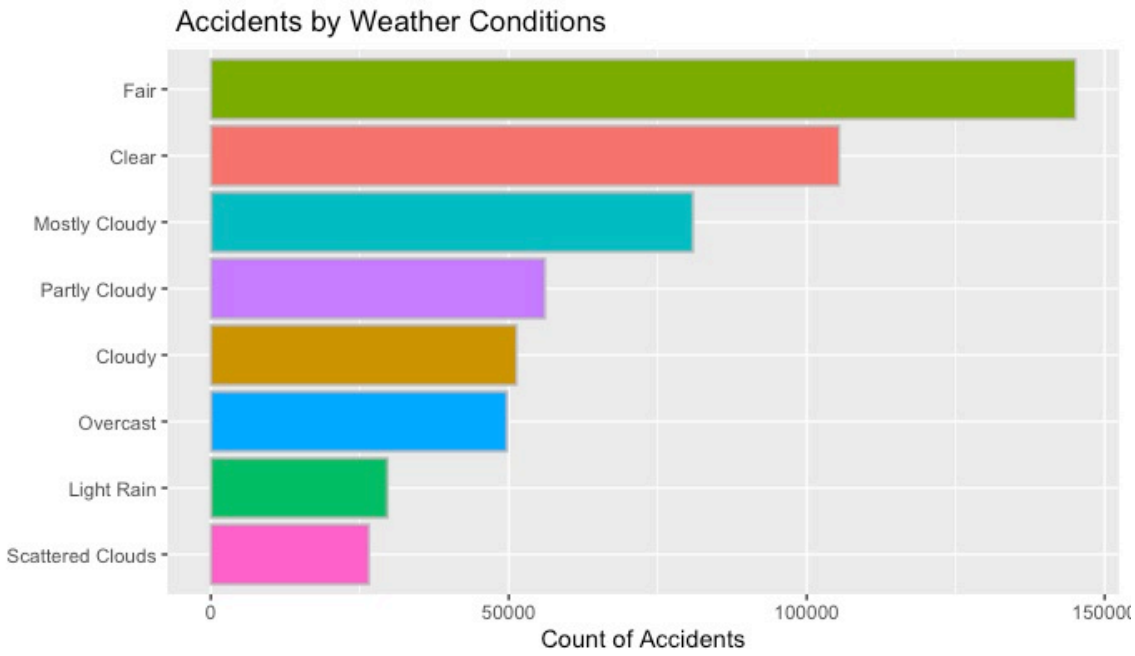


Figure 12: Accidents According to Weather Conditions

Figure 13 is the outcome of mutating the street column generating two outputs. As shown, most accidents occur in cities, not highways, due to the high density of population and cars in cities, proving the high impact of traffic on accidents.

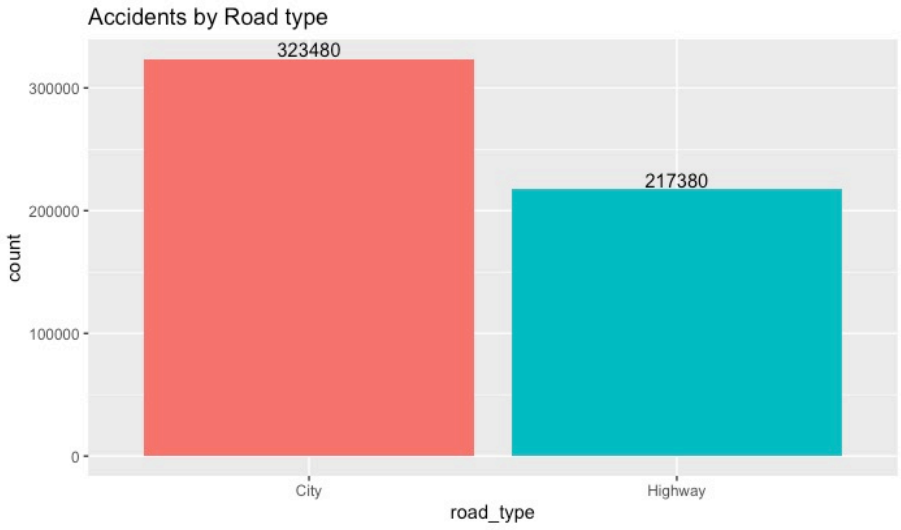


Figure 13: Accidents By Road Type

Cars drive faster on left lanes; however, *Figure 14* shows that most accidents occur on the right lane indicating that speed is not a significant contributor to car accidents. Car collisions can still happen even if drivers are not speeding or exceeding the speed limit.

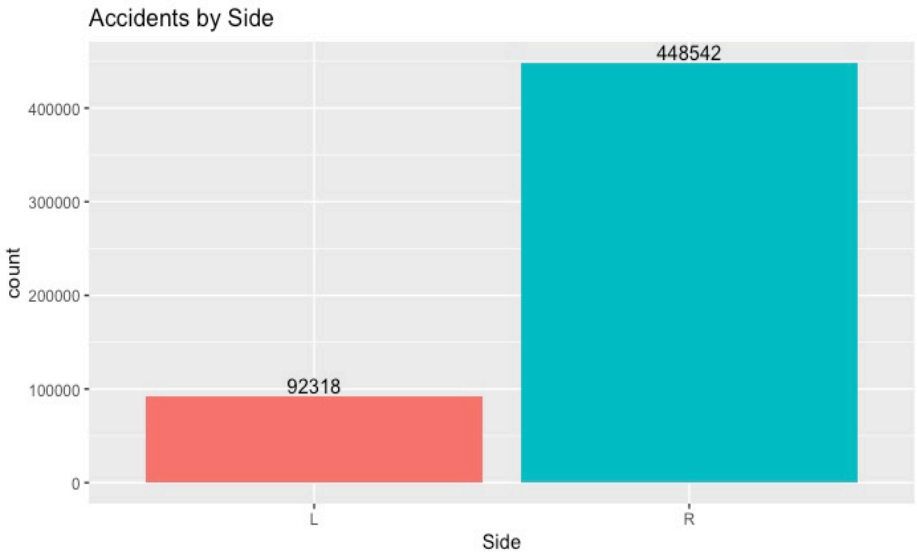


Figure 14: Accidents by Side

Figure 15 represents the availability of annotations in the area, first the column was filtered to results = TRUE, which means that those annotations were available when accidents took place. Results showed that a high number of accidents occur near traffic signals, junctions, and crossings indicating that drivers do not leave enough safety distance between their car and the car in front of them, in addition to lack of attention producing human errors. For example, when a yellow traffic light hits or when crossing a junction, causing impulse actions that often if not always leads to over speeding, or hitting an emergency brake resulting in a car collision.

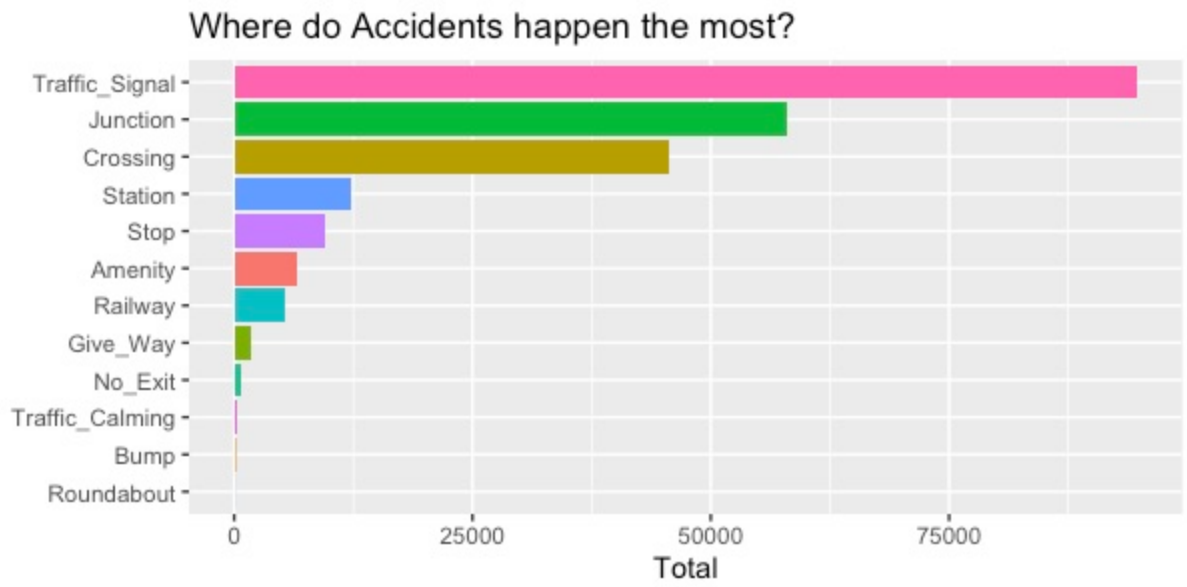


Figure 15: Accidents based on availability of Annotations

4.3.3 Analysis of Traffic Severity

Car accidents not only result in casualties, injuries, and property damage, but also produce traffic which may cause more accidents. The severity column is the target selected variable for the analysis.

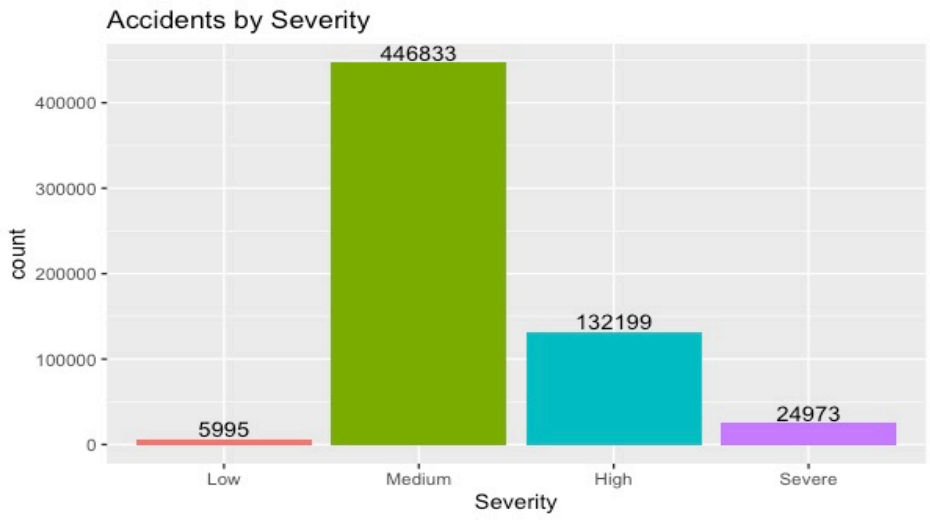


Figure 16: Traffic Severity Effect from the accidents

As *Figure 16* illustrates, the distribution is very imbalanced, therefore both “Low” and “Medium” were combined and classified as “Not Severe”. As well as “High” and “Severe” were combined and classified as “Severe”.

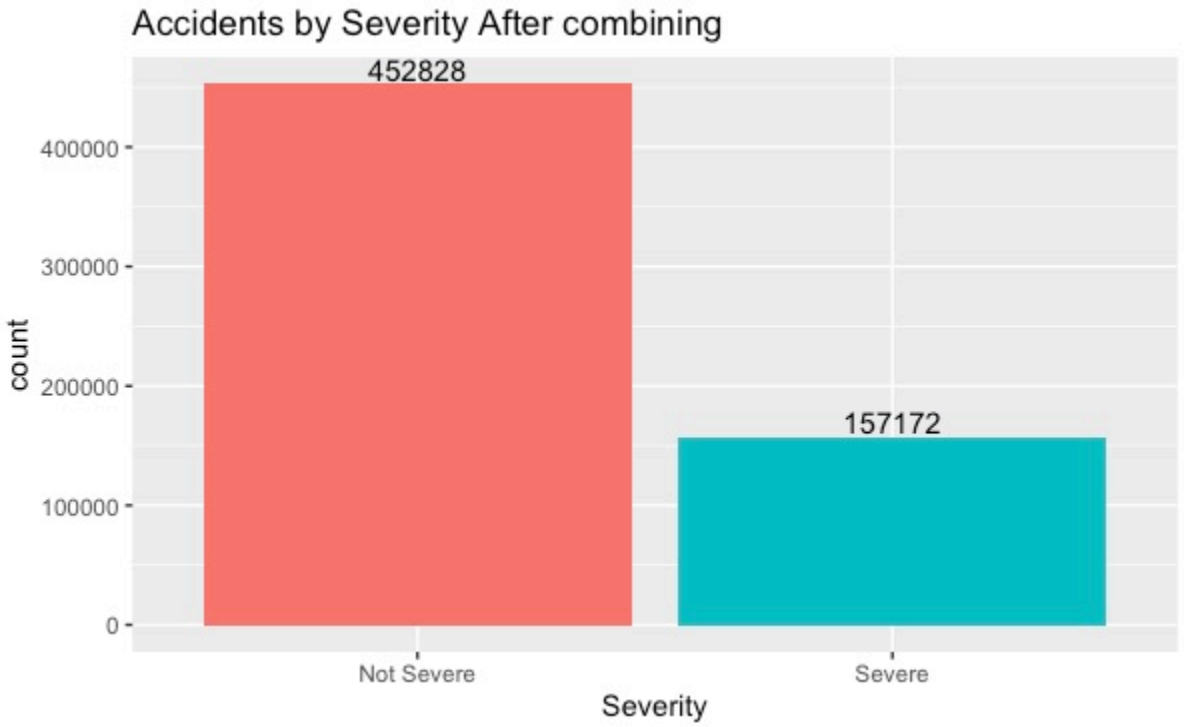


Figure 17: Modification of Traffic Severity column

Figure 17 illustrates the modification implemented After splitting the severity column into “Severe”, or “Not Severe” a correlation matrix was performed to see which variables have effect on Traffic severity, *Figure 18* illustrates the relationship between Severity and Annotations; results showed that there is no strong correlation between those variables. *Figure 19* uses the rest of the variables, results showed that “Road type” and “Side” has the strongest effect on Traffic Severity.

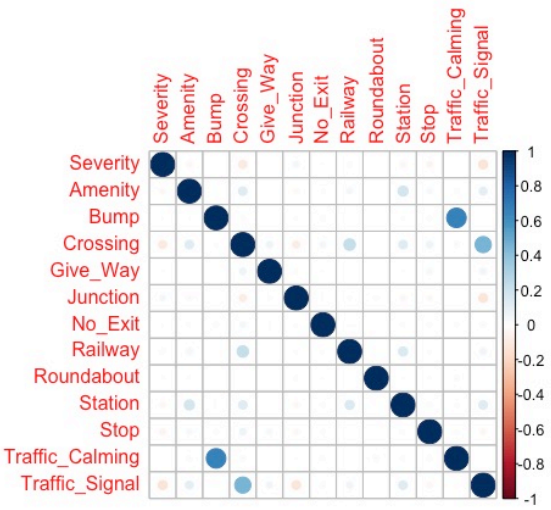


Figure 18: Correlation Matrix 1

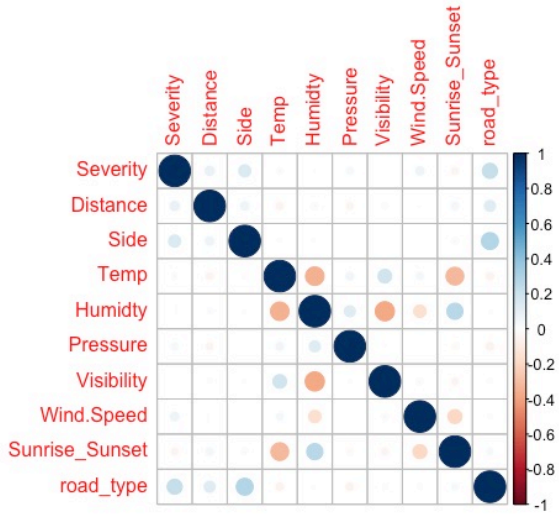


Figure 19: Correlation Matrix 2

As a result of the correlation matrix, a representation of the relationship between side, road type, and severity is illustrated

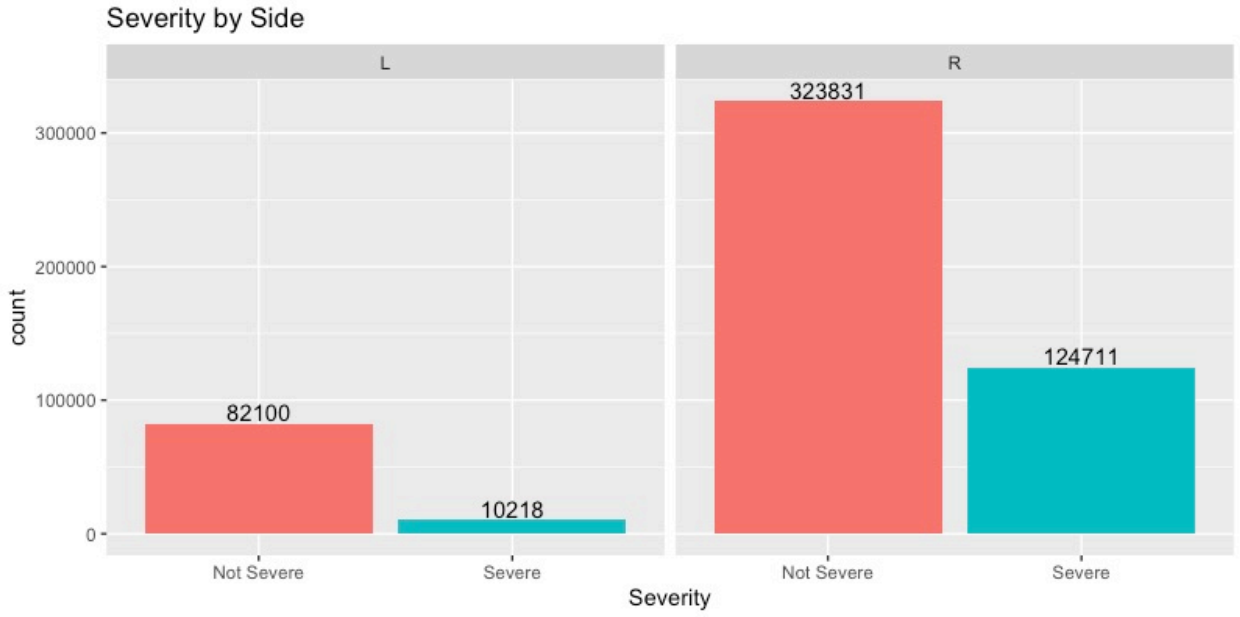


Figure 20: Traffic Severity by Side

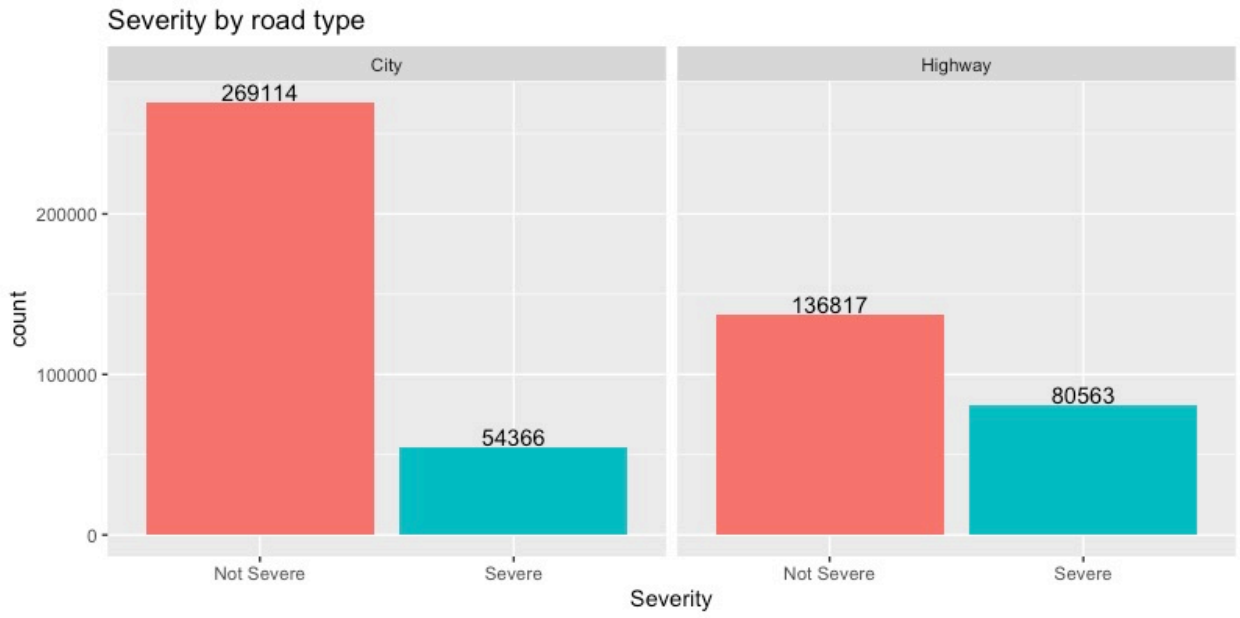


Figure 21: Traffic Severity by Road Type

Figure 21 shows that the majority of accidents that cause severe traffic occur at the right side; which corresponds with our finding back in *Figure 14.* However; *Figure 21* states that severe traffic accidents occur more on highways not cities; due to the fact that drivers speed more on highways compared to cities, in addition to the fact, highway accidents are very severe in terms of both injuries and impact on traffic.

4.4 Model Building

The last part consists of predicting the traffic severity of the accident. Two different supervised learning algorithms were chosen to be applied. The first one is the Naive Bayes which is a probabilistic machine learning model that uses the Bayes theorem. ***Bayes theorem*** is defined as finding the probability of "A" given that "B" has occurred. All predictors are independent; therefore, the presence of one predictor does not affect the other; hence, it's called Naive which is represented by the following equation: $P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$

The second algorithm chosen was the random forest which consists of a collection of uncorrelated decision trees, which are then merged together to reduce variance and create more accurate data predictions.However, looking at *Figure 17* we can see that the distribution of severity is imbalanced, therefore for better results, it should be balanced. *Figure 22* illustrates the severity column after balancing,

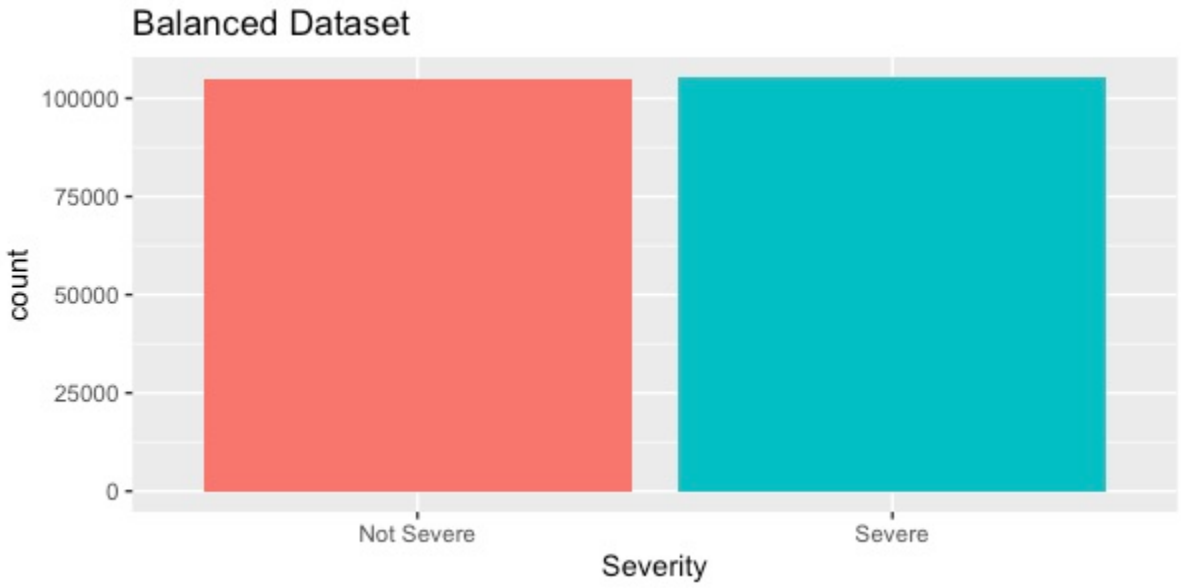


Figure 22: Traffic Severity After Balancing

4.5 Comparison of Different Models

Naive Bayes Results

	Actual	
Prediction	**Severe**	**Not severe**
Severe	233,209	72,668
Not Severe	37,141	198,011

Table 3: Naive Bayes Confusion Matrix

The confusion matrix represents the True positive, how many times the model predicted *Severe* correctly. True negative how many times the model predicted **Not Severe** correctly. False Positive how many times the model predicted *Severe* incorrectly. False negative how many times the model predicted **Not Severe** incorrectly. *Table 4* will include the statistics extracted from the confusion matrix.

	Accuracy	**Sensitivity**	**Specificity**	**Pos. pred. value**	**Neg. pred. value**	**Positive Class**
Results	79.7%	86.26%	73.15%	76.24%	84.21%	Severe
Formula	$\frac{TP+TN}{TP+TN+FP+FN}$	$\frac{TP}{TP+FN}$	$\frac{TN}{TN+FP}$	$\frac{TP}{TP+FP}$	$\frac{TN}{TN+FN}$	

Table 4: Naive Bayes Statistics

Table 4 illustrates the results from performing the Naive Bayes algorithm, the model has an accuracy of almost 80%, which is good. The model had a sensitivity of 86.26% which means the rate in determining the **Severe** cases correctly is high; however the specificity which is the rate in determining **Not Severe** cases was lower showing a successful rate of only 73.15%. The model

has a 76.24% positive predictive value, which is the case of predicting only **Severe** cases, and a negative predictive value of 84.21% which is the rate in predicting only **Not Severe** cases. Overall the model has a good accuracy rate, it has a good sensitivity ratio for predicting the **Severe** cases, however the specificity ratio is low.

Random Forest Results:

	Actual	
Prediction	**Severe**	**Not severe**
Severe	237,239	53,748
Not Severe	33,111	216,931

Table 5: Random Forest Confusion Matrix

	Accuracy	**Sensitivity**	**Specificity**	**Pos. pred. value**	**Neg. pred. value**	**Positive Class**
Results	84%	87.75%	80.14%	81.5%	86.76%	Severe

Table 6: Random Forest Statistics

As *Table 6* illustrates, the random forest algorithm outperformed the Naive Bayes, it resulted with a better accuracy of 84%, as well as better sensitivity ratio of 87.75%. Moreover, it outperformed the specificity of the naive bayes model, scoring 80.14% success rate in predicting **Not severe** cases compared to 73.15% success rate. This summarizes our findings, that yes it is possible to predict the **traffic severity** when having the following variables: time of the accident, day of the accident, location of the accident, the road type, the side, and the availability of any near annotation.

Chapter 5 - Conclusion

5.1 Conclusion

This study is meant to understand the significance of the contributors that cause car accidents and provide solutions to minimize them. Results showed that traffic is the primary cause of car accidents. The majority of accidents occur at rush hour when people or students are either going to work or school or coming back home.

Moreover, accidents tend to happen in cities with high population density. Different weather conditions, speed, and lightning had a low impact on accidents. Finally, attributes such as time, state, city, and street can help official authorities predict traffic severity in advance and provide alternative solutions ahead to reduce traffic and prevent road accidents.

5.2 Possible Solutions

Knowing that traffic is the major contributor to road accidents, reducing traffic will eventually reduce accidents. Many solutions can be taken into consideration. One of them is adopting the work from home strategy; covid has proved that many employees can finish their tasks from home without physically being in the office. This will reduce the heavy load on roads and reduce traffic. Another possible solution is to set apart the starting time of schools and working organizations; this will help divide the number of cars on streets among multiple hours.

The final solution is to encourage and facilitate the adoption of self-driving vehicles since accidents mainly occur because of human errors due to lack of attention, concentration, and making wrong decisions. Using automated cars will resolve this issue by reducing human errors, which will reduce car accidents.

5.3 Future Work

Future studies can involve analyzing different countries, and see if they will produce similar results, as well as more details about the driver and the vehicle. Information such as age, gender, profession, car type, and ownership can be valuable. This will help to analyze the psychological behavior of the individuals and understand how those aspects might affect the individual driving behavior.

Bibliography (APA Format)

National Center for Statistics and Analysis. (2020, December). Overview of motor vehicle crash- es in 2019. (Traffic Safety Facts Research Note. Report No. DOT HS 813 060). National High- way Traffic Safety Administration. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813060

National Center for Statistics and Analysis. (2020, March). Pedestrians: 2018 data (Traf- fic Safety Facts. Report No. DOT HS 812 850). National Highway Traffic Safety Administration. Retrieved from; https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812850

National Center for Statistics and Analysis. (2019, December). Seat belt use in 2019 – Overall Results (Traffic Safety Facts Research Note. Report No. DOT HS 812 875). National Highway Traffic Safety Administration. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812875

National Center for Statistics and Analysis. (2020, October). Preview of motor vehicle traffic fatalities in 2019 (Research Note. Report No. DOT HS 813 021). National Highway Traf- fic Safety Administration. Retrieved from: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813021

Horne, J., & Reyner, L. (1999). Vehicle accidents related to sleep: A review. Occupational and Environmental Medicine. BMJ Publishing Group. https://doi.org/10.1136/oe_m.56.5.289

Farooq, D., Juhasz, J. (2020). "Simulation Analysis of Contributing Factors to Rider Visibility Issues for Car-Motorcycle Accidents", Periodica Polytechnica Transportation Engineering, 48(3), pp. 203–209. https://doi.org/10.3311/PPtr.13521

Fan, F. (2018). Study on the Cause of Car Accidents at Intersections. Open Access Library Journal, 5: e4578. https://doi.org/10.4236/oalib.1104578

Chen, C., Zhao, X., Liu, H., Ren, G., & Liu, X. (2019). Influence of adverse weather on drivers' perceived risk during car following based on driving simulations. Journal of Modern Transportation, 27(4), 282–292. https://doi.org/10.1007/s40534-019-00197-4

Mao, X., Yuan, C., Gan, J., & Zhang, S. (2019). Risk factors affecting traffic accidents at urban risk factors affecting traffic accidents at urban weaving sections: Evidence from China. International Journal of Environmental Research and Public Health,16(9). https://doi.org/10.339_0/ijerph16091542

Jägerbrand, A. K., & Sjöbergh, J. (2016). Effects of weather conditions, light conditions, and road lighting on vehicle speed. SpringerPlus, 5(1). https://doi.org/10.1186 / s40064-016-2124-6

Jadayil, W. A., Khraisat, W., & Shakoor, M. (2020). Statistical analysis for the main fac- tors causing car accidents. ARPN Journal of Engineering and Applied Sciences, 15(5), 696–715. Retrieved from: http://www.arpnjournals.org/jeas/research_papers/rp_2020/jeas_0320_8150.pdf

Miller, T. R., Bhattacharya, S., Zaloshnja, E., Taylor, D., Bahar, G., & David, I. (2011). Costs of crashes to Government, United States, 2008. In Annals of Advances in Automotive Med- icine (Vol. 55, pp. 347–355). Retrieved from: https://www.ncbi.nlm.nih.gov/ pmc/ articles /PMC3256 813/

Tadege, M. (2020). Determinants of fatal car accident risk in Finote Selam town, Northwest Ethiopia. BMC Public Health 20, 624. https://doi.org/10.1186/s12889-020-08760-z

Celik AK, Oktay E. (2014). A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars provinces of Turkey. Accident Analysis prev. 72:66–77. Retrieved from: https://doi.org/10.1016/j.aap.2014.06.010

Gicquel L, Ordonneau P, Blot E, Toillon C, Ingrand P and Romo L (2017). Description of Various Factors Contributing to Traffic Accidents in Youth and Measures Proposed to Alleviate Recurrence. Front. Psychiatry 8:94. https://doi.org/10.3389/fpsyt.2017.00094

CépiDC. Centre d'épidémiologie sur les causes médicales de décès. (2015). Retrieved from: http://www.cepidc.inserm.fr/site4/

Waylen AE, McKenna FP. (2008). Risky attitudes towards road use in pre-drivers. Accident Analysis Prev 40(3):905–11. https://doi.org/10.1016/j.aap.2007.10.005

Saifuzzaman M, Haque M, Zheng Z, Washington S. (2015). Impact of mobile phone use on car-following behavior of young drivers. Accident Analysis Prev 82:10–9. https://doi.org/10.1016/j.aap.2015.05.001

Knipling RR, Wang J-S. (1994). Crashes and fatalities related to driver drowsiness/fatigue. Washington, DC, OYce of Crash Avoid- ance Research, US Department of Transportation, Research note. Retrieved from: https://rosap.ntl.bts.gov/view/dot/2936/dot_2936_DS1.pdf

Boyce PR (2003) Human factors in lighting, 2nd edn. Taylor & Francis, London Brodsky H, Hakkert AS (1988) Risk of a road accident in rainy weather. Accident Analysis Prev 20(3):161–176. https://doi.org/10.1016/0001-4575(88)90001-2

Michel Bédard, Gordon H. Guyatt, Michael J. Stones, John P. Hirdes. (2002). The independent contribution of driver, crash, and vehicle characteristics to driver fatalities, Accident Analysis & Prevention, Volume 34, Issue 6, Pages 717-727, https://doi.org/10.1016/S0001-4575(01)00072-0.