Road accidents analysis and prediction

On US Accidents

Nassima BENAMMAR
22/09/2020

Table des matières

1. Introduction

1.1. Context

Road traffic is a part of everyone's daily life since cars has become accessible in all over the world. Unfortunately, the number of car accidents is much greater than any other mean of transport. Between 2008 and 2010, in all the territories of the 27 EU countries, the number of deaths per billion kilometres was 3.14 passengers for the single car, 0.20 for bus passengers, 0.13 for trains and 0.06 for planes (European Union Agency for Railways). With the exponential evolution of Data science, we can provide a relevant data analysis on Road traffic accidents in order to prevent from fatal accidents in the future.

A road accident is related to several factors such as weather, type of the road, time and place. A data analysis will provide a very useful overview on the road traffic accidents: what is corelated with the severity of an accident, what type of road is mostly cited in accident descriptions, what location occurs the most, etc. Plus, if we learn enough data, we can build a relevant predictive model. In reality, it is difficult to predict if there will be an accident or not given a set of weather, place or time values. Nevertheless, we can predict the severity of an accident if there is any with those values.

1.2. Interest

For our study, we will focus on the USA road accidents. We will first describe the chosen data sets we selected for this study and we will present a statistical analysis for a better comprehension of the dataset. After that, we will prepare the data in order to build a relevant predictive model. We will use several algorithms of machine learning and compare them. The main objective is to produce accurate prediction of road traffic accidents in the USA.

2. Data

2.1. Data source

For this study, we will use the US accidents dataset from https://smoosavi.org/datasets/us_accidents which represents a countrywide traffic accident dataset over 49 states of the USA from 2016 until June 2020. Today, there are about 3 ,5 million accidents recorded in the data set. These data are collected from several data providers that capture streaming traffic event from state departments of transportation, law enforcement agencies, traffic cameras and traffic sensors.

There are 3513617 records and 49 columns in the dataset. The attributes can be categorized as follows:

General description: all the attributes are not nullable except TMC and longitude and latitude of the end point (see the descriptions below)

ID	identifier of the accident. 3513617 unique values
Source	the source of the record. 3 unique values
TMC	more description of the event. 21 unique values
Severity	the severity of the accident (from 1 to 4 for fatal accident) which the target.
Start_time, End_time	the start time and the end time of the accident

Star_lat, star_lng	le latitude and the longitude of the start point (in GPS)
End_lat, End_lng	le latitude and the longitude of the end point (in GPS)
Distance	the length in miles of the road where the accident happened
Description	the description has 1780092 values.

Address: all the attributes are nullable

Number	the street number
Street	the name of the street
Side	the relative side of the street (R/L).
City	the city where the accident happened.
County	the county of the accident.
State	the state of the accident (49 states)
Zip code	the zip code of the accident
Country	the country of the accident (one country).
Timezone	based on the location of the accident (eastern, central, etc.).
Airport_code	the closest airport (relevant for rental cars).

Weather: all the attributes are nullable

Weather_timestamp	the time-stamp of the weather observation record (local time).
Temperature, wind_chill	in Fahrenheit.
Humidity	in percentage.
Pressure	the air pressure in inches.
Visibility	in miles
Wind_direction	wind direction (24 values).
Wind_speed	in miles per hour.
Precipitation	precipitation amount in inches if there is any.
Weather_Cond	the weather condition (rain, snow, clear, etc.) with 127 values.

POI Annotation: these attributes are Boolean and are not nullable.

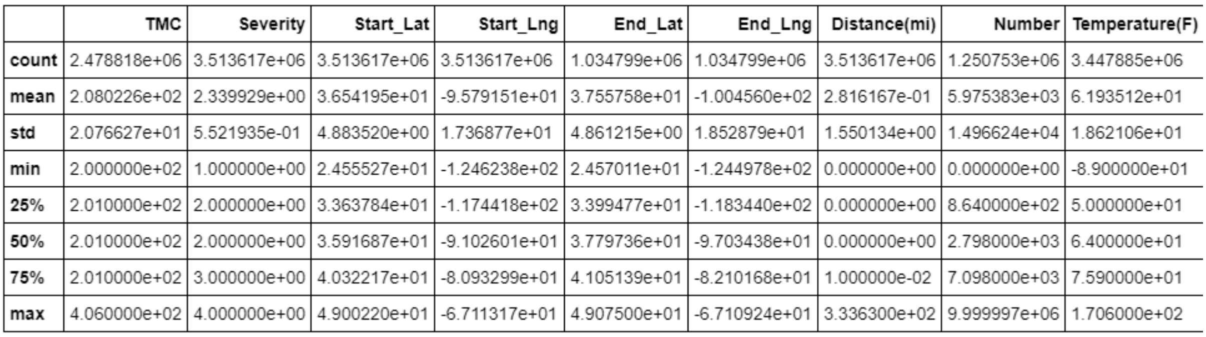
Amenity	amenity in a nearby location.
Bump	speed bump or hump in a nearby location.
Crossing	crossing in a nearby location.
Give_Way	give_way in a nearby location.
Junction	junction in a nearby location.
No_Exit	no_exit in a nearby location.
Railway	railway in a nearby location.
Roundabout	roundabout in a nearby location.
Station	station in a nearby location.
Stop	stop in a nearby location.

Traffic_Calming	traffic_calming in a nearby location.
Traffic_Signal	traffic_signal in a nearby location.
Turning_Loop	turning_loop in a nearby location.

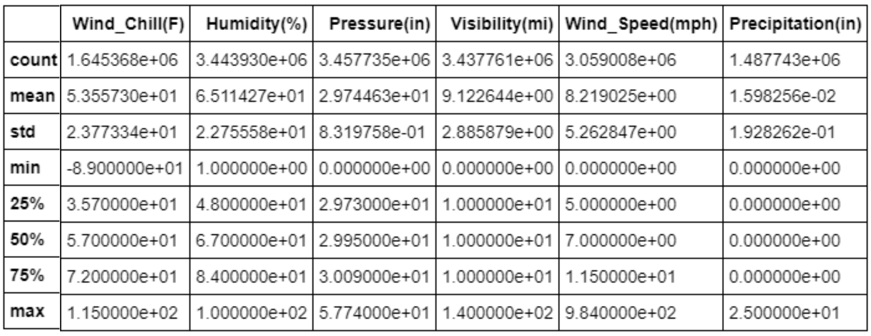
Period of the day : the attributes can be nullable.

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

To have a better idea on the numeric attributes, we display here the statistics given by the function “describe”:



	TMC	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Number	Temperature(F)
count	2.478818e+06	3.513617e+06	3.513617e+06	3.513617e+06	1.034799e+06	1.034799e+06	3.513617e+06	1.250753e+06	3.447885e+06
mean	2.080226e+02	2.339929e+00	3.654195e+01	-9.579151e+01	3.755758e+01	-1.004560e+02	2.816167e-01	5.975383e+03	6.193512e+01
std	2.076627e+01	5.521935e-01	4.883520e+00	1.736877e+01	4.861215e+00	1.852879e+01	1.550134e+00	1.496624e+04	1.862106e+01
min	2.000000e+02	1.000000e+00	2.455527e+01	-1.246238e+02	2.457011e+01	-1.244978e+02	0.000000e+00	0.000000e+00	-8.900000e+01
25%	2.010000e+02	2.000000e+00	3.363784e+01	-1.174418e+02	3.399477e+01	-1.183440e+02	0.000000e+00	8.640000e+02	5.000000e+01
50%	2.010000e+02	2.000000e+00	3.591687e+01	-9.102601e+01	3.779736e+01	-9.703438e+01	0.000000e+00	2.798000e+03	6.400000e+01
75%	2.010000e+02	3.000000e+00	4.032217e+01	-8.093299e+01	4.105139e+01	-8.210168e+01	1.000000e-02	7.098000e+03	7.590000e+01
max	4.060000e+02	4.000000e+00	4.900220e+01	-6.711317e+01	4.907500e+01	-6.710924e+01	3.336300e+02	9.999997e+06	1.706000e+02



	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
count	1.645368e+06	3.443930e+06	3.457735e+06	3.437761e+06	3.059008e+06	1.487743e+06
mean	5.355730e+01	6.511427e+01	2.974463e+01	9.122644e+00	8.219025e+00	1.598256e-02
std	2.377334e+01	2.275558e+01	8.319758e-01	2.885879e+00	5.262847e+00	1.928262e-01
min	-8.900000e+01	1.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00
25%	3.570000e+01	4.800000e+01	2.973000e+01	1.000000e+01	5.000000e+00	0.000000e+00
50%	5.700000e+01	6.700000e+01	2.995000e+01	1.000000e+01	7.000000e+00	0.000000e+00
75%	7.200000e+01	8.400000e+01	3.009000e+01	1.000000e+01	1.150000e+01	0.000000e+00
max	1.150000e+02	1.000000e+02	5.774000e+01	1.400000e+02	9.840000e+02	2.500000e+01

Let’s focus on the Severity, Distance and Weather field. For the severity, the average is about 2.3 which means that in general the road accidents are not extremely severe. 25% of the accidents have a severity greater than 2. The distance of an accident is generally close to 0,3 mile with a maximum of 334 mile which is not normal.

For the weather field, the temperature is generally about 60°F. It is difficult to be accurate in our analysis here because the USA is almost a continent with different climate characteristics from north to south and from east to west. However, we can have a general idea on the average visibility during an accident which is 9 miles. It represents a small visibility but, in a road, it is quite sufficient to have a good visibility in the horizon. So, it doesn’t represent obviously a bad weather. Same thing for Wind speed and Pressure. The average Humidity is 65% but with a std of 10% which means that the average is weak. Furthermore, 75% of accidents occurred while the humidity was less than 72%.

2.2. Dealing with missing values

In this dataset well synthetised by Smoosavi, there are 13061803 missing values. Among all the attributes, 23 have at least one missing value. We need to figure out with this issue.

Attribute	Number of NaN values
TMC	1034799
End_Lat	2478818
End_Lng	2478818
Description	1
Number	2262864
City	112
Zipcode	1069
Timezone	3880
Airport_Code	6758
Weather_Timestamp	43323
Temperature(F)	65732
Wind_Chill(F)	1868249
Humidity(%)	69687
Pressure(in)	55882
Visibility(mi)	75856
Wind_Direction	58874
Wind_Speed(mph)	454609
Precipitation(in)	2025874
Weather_Condition	76138
Sunrise_Sunset	115
Civil_Twilight	115
Nautical_Twilight	115
Astronomical_Twilight	115

We need to replace NaN values with other values. For the attributes with numeric values we will take the mean value. For categorical values such as Description, Timezone, City, Wind_Direction, Weather_condition and period of the field, we will process differently. For some of them we will take the mode value.

For categorical data, we will proceed as follows for each column :

- Description: there is only one missing value so we replace with the mode because it won't affect the result later
- City : 112 values represents 0.003 % of the values so the mode can replace the missing values without affecting the results.
- Weather_Timestamp: we can replace the missing values with the values of Start_Time of the accident which cannot be null. First, we need to convert Start_Time End_Time and Weather_Timestamp to datetime format.
- Weather_Condition : this attribute has almost 100 unique value. The mode can be sufficient.
- Wind_Direction : we first normalise the data by replacing the values North, West, East and South by N, W, E and S then we replace the NaN values with the mode.

2.3. Drop attributes

Some of the attributes are not necessarily relevant for our study such as TMC that is an additional description of the accident. I think that the other attributes of the general description are enough. Plus, it has lots of missing values. For example, almost 70% values of the End_Lat, End_Lng and Number attributes are missing. We thus prefer drop these attributes from the dataset.

In the address field, we drop the zip code since it has lots of missing values and we already have a well detailed address with the street, the city, the county and the state. We have also the latitude and the longitude of the start point which gives an idea on the exact address. We drop also the airport code because I don't think it is relevant for our study. Since the dataset is about the USA accidents road, it is not necessary to keep the Country attribute.

For Boolean attributes, which represent all the annotations attributes, it is better to convert them into integers (0 or 1).

For the other such as for the period of the day field, we only keep Sunrise_Sunset because the order attributes are not relevant as you can see bellow : (they represent additional measure of the period of the day).

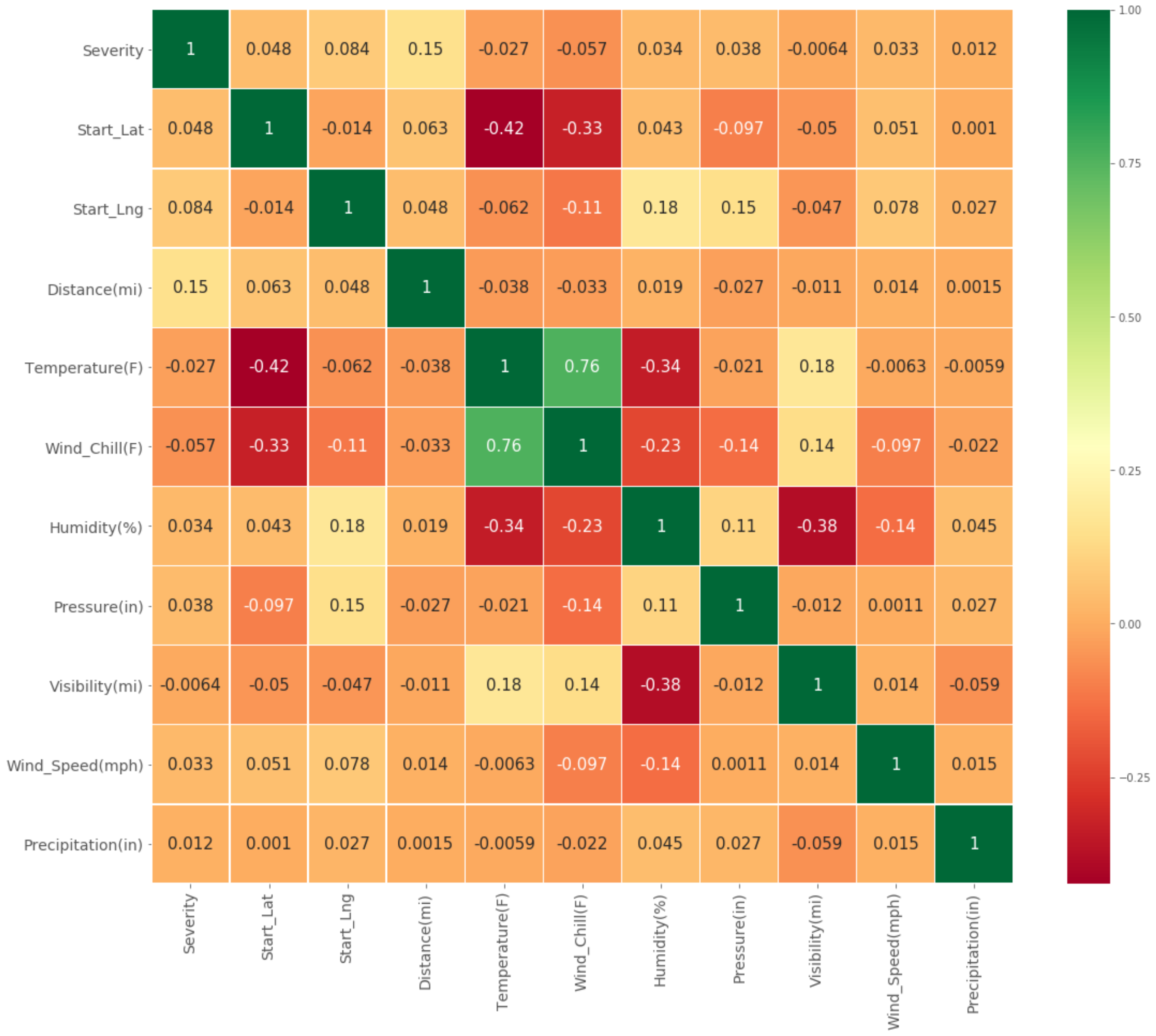
The following attributes are the final columns of our dataset : 'ID', 'Source', ‘Severity', 'Start_Time', 'End_Time', 'Start_Lat', 'Start_Lng', 'Distance(mi)', 'Description', 'Street', 'Side', 'City', 'County', 'State', 'Timezone', 'Weather_Timestamp', 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Direction', 'Wind_Speed(mph)', 'Precipitation(in)', 'Weather_Condition', 'Amenity', 'Bump', 'Crossing', 'Give_Way', 'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Station', 'Stop', 'Traffic_Calming', 'Traffic_Signal', 'Turning_Loop', 'Sunrise_Sunset'

3. Data analysis

In the data analysis, we will show what attributes are correlated or not with Severity and analyse the proportion of the severity of accidents under weather condition, depending on the place and the time in order to show to the people what can really impact the severity of an accident and hopefully prevent them from severe accidents.

3.1. Exploratory Data

We display the correlation grid between numeric attributes and the target which is Severity, no attribute has a significant correlation with it. It is than hard to predict the severity on the basis of these attributes. The maximum that we obtain is 0,057 with Wind_chill and it is a very weak correlation with the Severity.



	Severity	Start_Lat	Start_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
Severity	1	0.048	0.084	0.15	-0.027	-0.057	0.034	0.038	-0.0064	0.033	0.012
Start_Lat	0.048	1	-0.014	0.063	-0.42	-0.33	0.043	-0.097	-0.05	0.051	0.001
Start_Lng	0.084	-0.014	1	0.048	-0.062	-0.11	0.18	0.15	-0.047	0.078	0.027
Distance(mi)	0.15	0.063	0.048	1	-0.038	-0.033	0.019	-0.027	-0.011	0.014	0.0015
Temperature(F)	-0.027	-0.42	-0.062	-0.038	1	0.76	-0.34	-0.021	0.18	-0.0063	-0.0059
Wind_Chill(F)	-0.057	-0.33	-0.11	-0.033	0.76	1	-0.23	-0.14	0.14	-0.097	-0.022
Humidity(%)	0.034	0.043	0.18	0.019	-0.34	-0.23	1	0.11	-0.38	-0.14	0.045
Pressure(in)	0.038	-0.097	0.15	-0.027	-0.021	-0.14	0.11	1	-0.012	0.0011	0.027
Visibility(mi)	-0.0064	-0.05	-0.047	-0.011	0.18	0.14	-0.38	-0.012	1	0.014	-0.059
Wind_Speed(mph)	0.033	0.051	0.078	0.014	-0.0063	-0.097	-0.14	0.0011	0.014	1	0.015
Precipitation(in)	0.012	0.001	0.027	0.0015	-0.0059	-0.022	0.045	0.027	-0.059	0.015	1

Figure 1Correlation between severity and quantitative features

3.2. Data analysis

We can at least give a data analysis on the time, the weather and the place that impact the severity of a road accident in the USA.

3.2.a. Weather data analysis

Most of the accident occurs in a good weather condition (33,5% in clear weather and 20.5% in fair weather), on the contrary of what we expect. It means that people drive carefully in bad weather condition which reduces the road accidents.

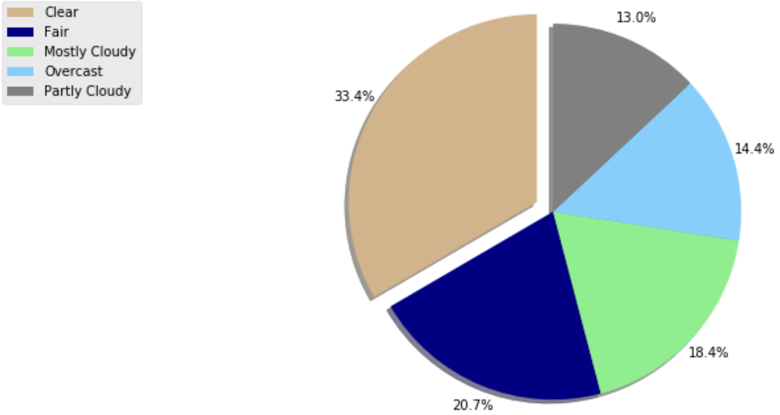


Figure 2 Number of accidents according to the weather condition

The following plots show that the more the temperature is greater the less is the severity of the accidents. And for humidity level of 50% , the severity is low. The pictures show also that the visibility and the severity have no impact on the severity level.

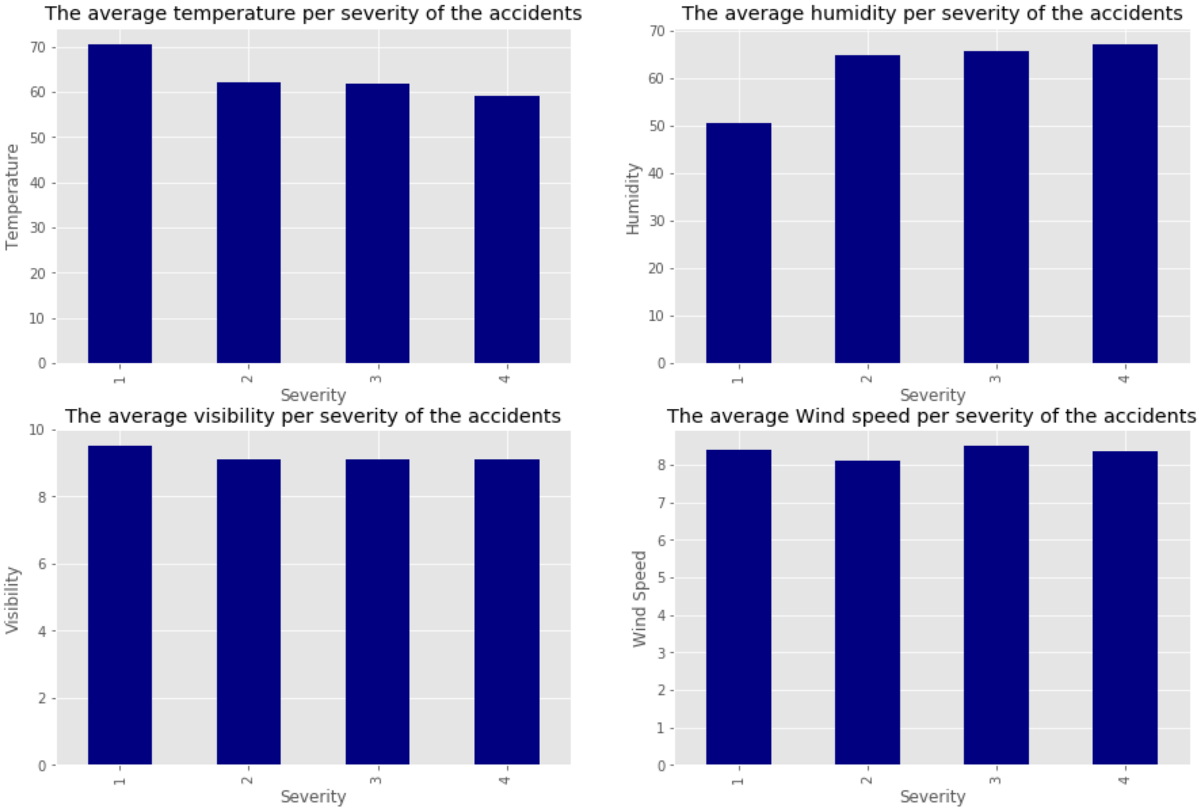


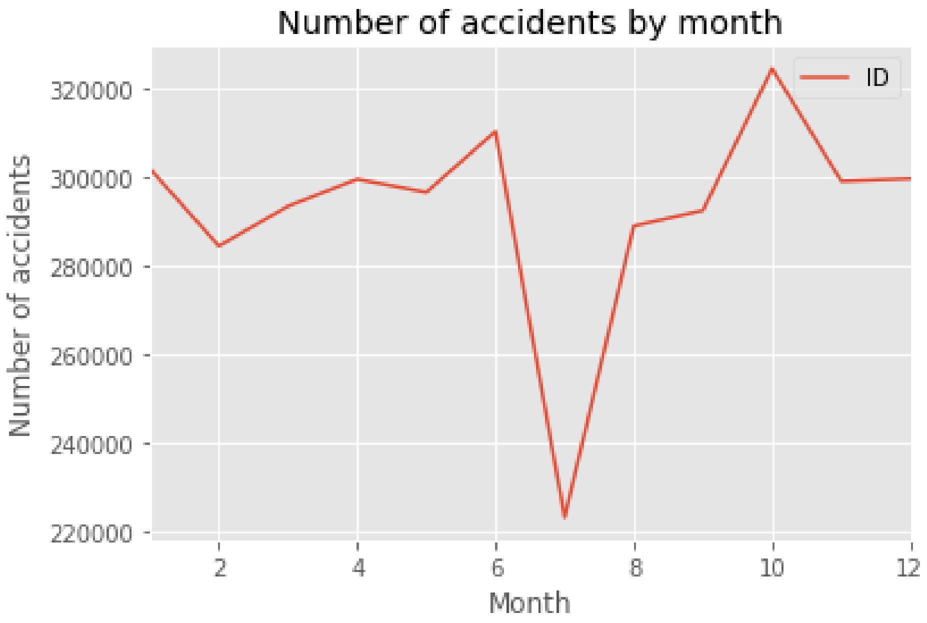
Figure 3 Impact of Temperature, Humidity, Visibility and Wind speed on the severity level

3.2.b. Time data analysis

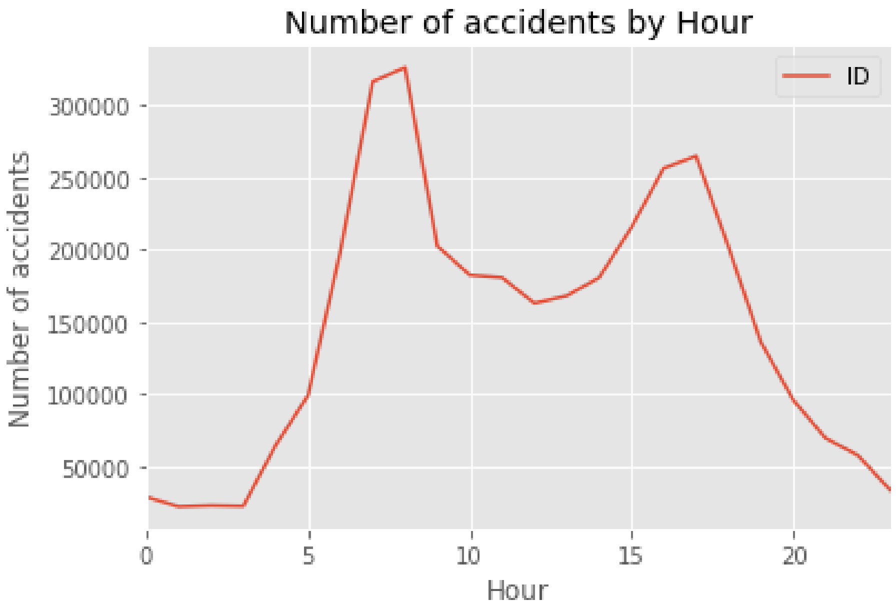
In the time analysis, we first displayed the evolution of the number of accidents from 2016 and 2020 which shows an increasement of the number of accidents until the end of 2019.

In 2020, the number of accidents crashed to the half which is probably due to the lock down because of the Covid 2019. Since the Watson IBM Lite was over, we couldn't get back all the figures.

We have at least the following pictures which show the number of accidents by month and by hour. It shows that the greatest number of accidents occurs around October before decreasing in Christmas holidays. Then it increases after the winter until June where it crashes during the summer holidays. This means that the seasons and the people activities are real factors on the number of accidents.



As expected, the number of accidents reaches the maximum in peak hours. It incredibly decreases by night. As for bad weather, people drive more carefully in bad condition so they also drive carefully by night.



3.2.c. Place data analysis

First, we show bar charts about the state and cities with the greatest number of accidents. By the following chart, we can see clearly that the number of accidents is much higher in California than in any other state. It is followed by Texas and Florida. This is normal because California is one the most crowded states with at least two huge cities.

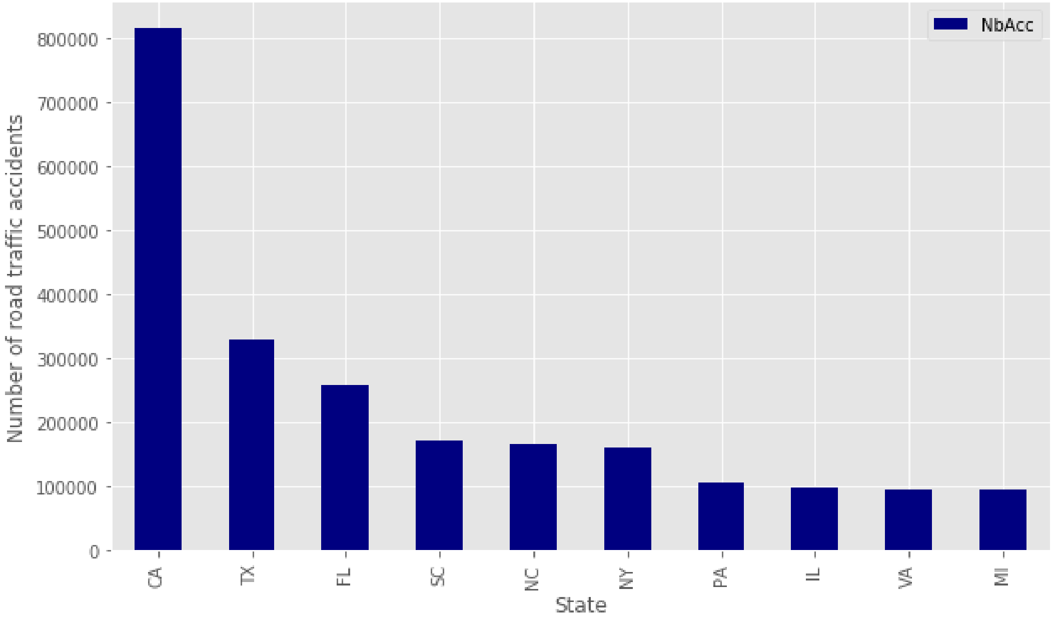
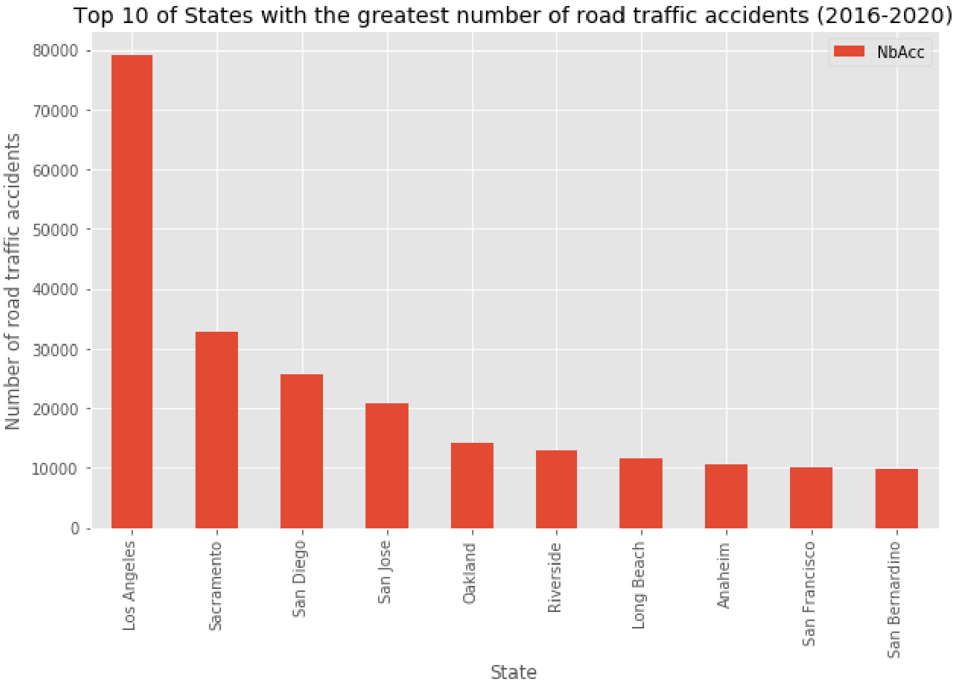


Figure 4 Top 10 of states with the greatest number of road traffic accidents (2016-2020)

We thus focus here on California to zoom on the cities that increase the number of accidents that much. We can see that Los Angeles is the cluster of danger in the road traffic accident. It is followed by Sacramento which has less than the half of the number of accidents in LA. The other cities are not far from LA such as Oakland and Long Beach. San Francisco is a big city by does not have an important number of accidents compared to LA and Sacramento.



It is better to display place analysis using a Choropleth. This is why we add some pictures below :

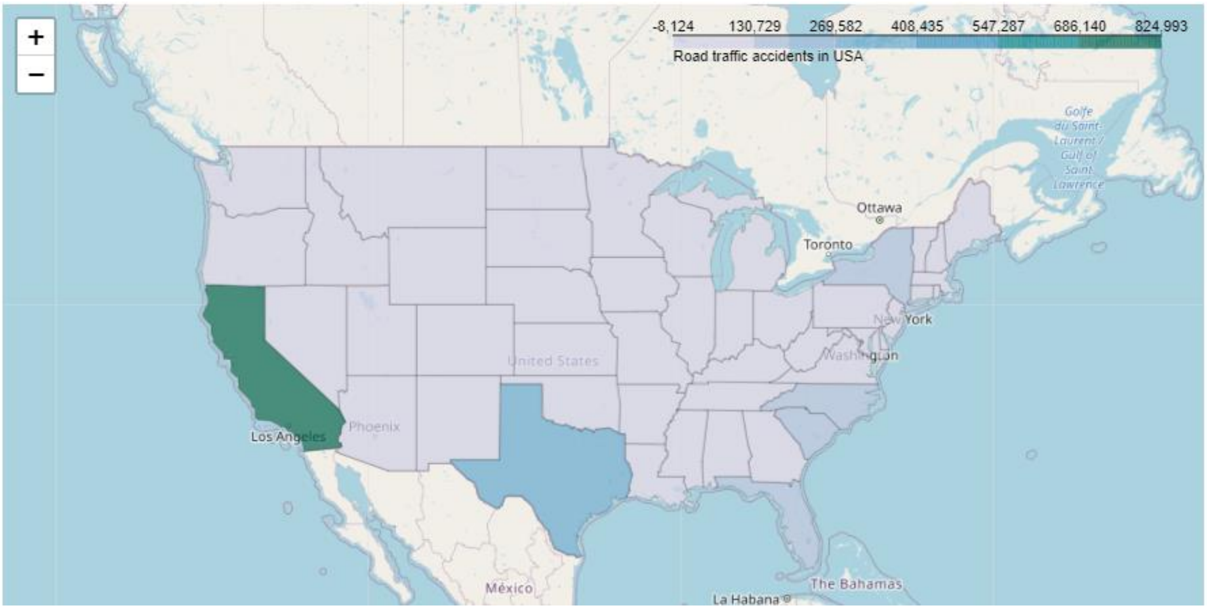


Figure 5 The most impacted state by road traffic accidents

California on top, which confirms the results in Figure 4.

Now let’s see what are the states where the accidents have a low severity (Severity=1). The difference between Figure 5 and Figure 6 is that Arizona is on top with California when the severity of the road accidents is low.

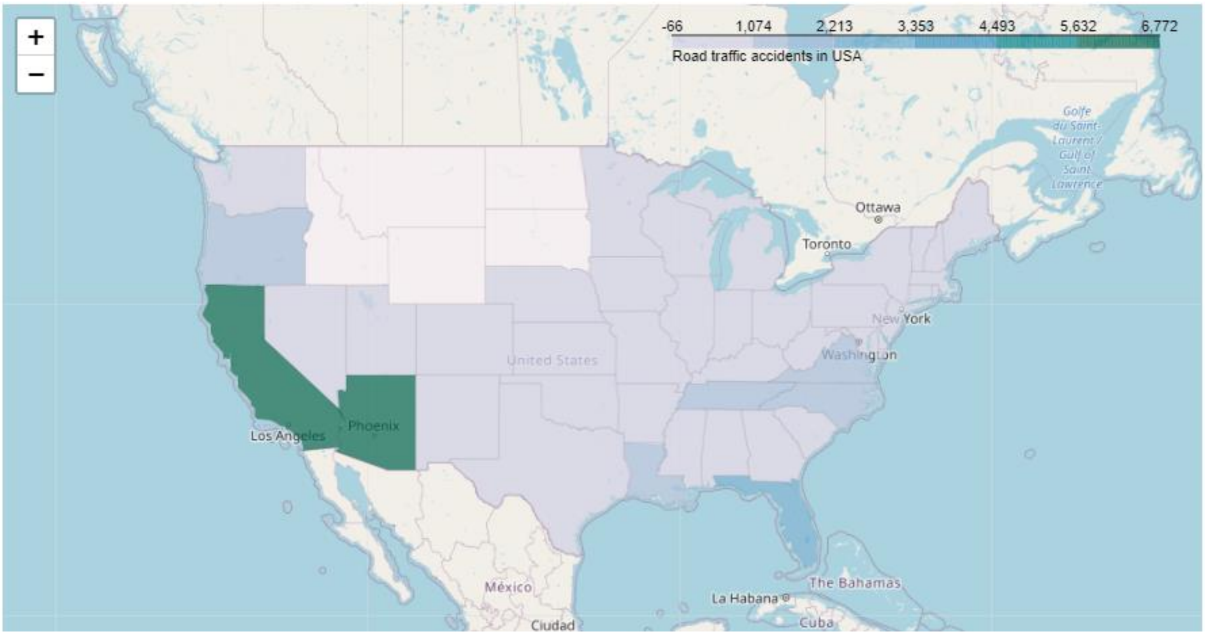


Figure 6 Number of road accidents by state with Severity=1

For very high severity (Severity=4), accidents occur in the states where the population is dense which means East cost and West Coast, plus Texas. We can also site Colorado because of Las Vegas city, Washington because of Seattle and Illinois because of Chicago.

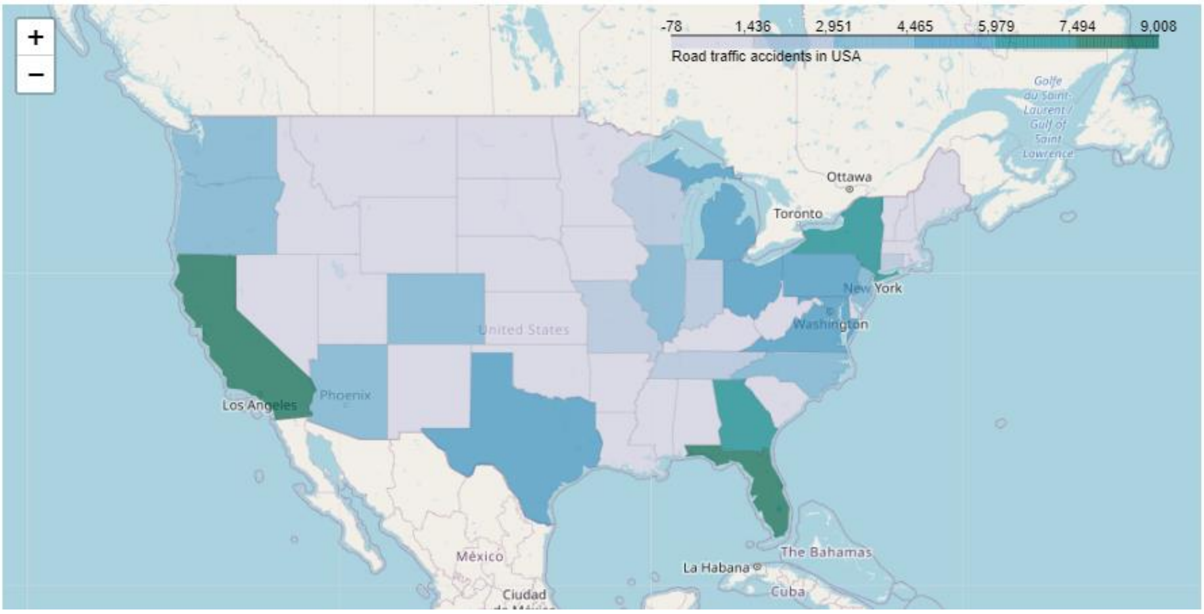


Figure 7 Number of road accidents by state with Severity=4

In the place field, one sub field is supposed to have a real impact on road accident. We can explore the traffic annotation field to determine what are the most traffic annotation that impact on the severity of an accident. Figure 8 depicts the number of accidents around traffic annotations

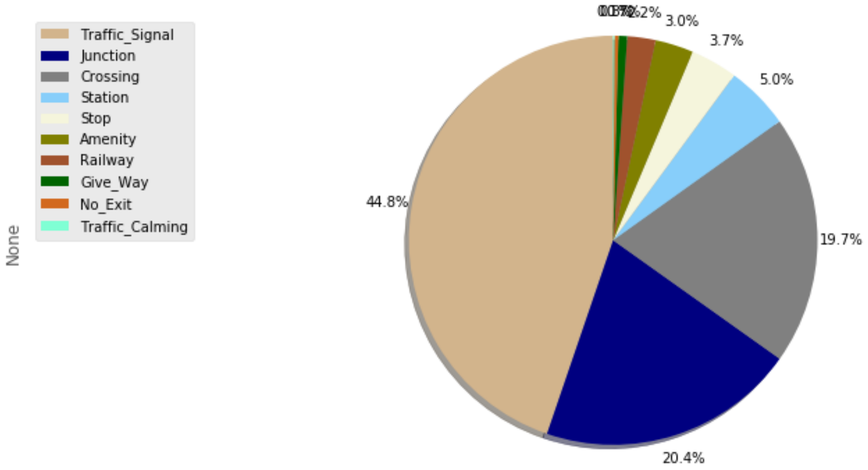


Figure 8 Number of accidents around traffic annotations

Almost half of the accidents occur around traffic signals. This is probably due to the people who rush when crossing the junctions while the red light just after the range light. Furthermore, 20% of the accidents occur in a junction and crossing places which are correlated to the traffic signal.

4. Discussion

4.1. Impact of features on the severity

The data set includes features about the place where the accident occurs, the time and the weather conditions. All these field allowed us to analyse the data and explain the relationship between the number of accidents and time, place and weather, and between the severity and those features.

We saw that some of them have a real impact on the number of accidents but not on the severity of accidents such for weather condition and time. It would be more relevant to add other features such as a description of the car or the driver and its state (speed, alcohol, crowded, stress, etc). Plus, we saw that most of the severe accident occur in crowded places but this conclusion is made on our general culture. We know that California is much more crowded than Montana. Consequently, the analysis would be more accurate is the machine had that same general culture represented by additional columns.

4.2. Predictive model

The dataset chosen is rich of information. It records different features for one accident. In data science, we can explore these features by statistical means and visualization to learn the phenomena of severe accidents. The case of the USA is really complex because the country is almost a continent and some states are completely different in climate, number of population and road complexity. The complexity of the country has a real impact on the correlation between the severity and other features. Is it relevant to build a predictive model with those means? Maybe we could try to have an idea and test with cross validation to have relevant results.

We saw in our study that the weather does not impact the severity of the accident in Figure 3. We only know that most of accident occur in good weather condition. We also saw that the most severe accident occurs in the states where the population is dense. Furthermore, most accidents occur around Traffic signals, junctions and crossing which are also correlated to the density of the population. Unfortunately, the dataset does not include this factor otherwise it would be really interesting to build a predictive model with the features :

- Weather condition
- Traffic signals
- Junctions
- Crossing
- Hour
- Month
- Population (not included in the data set)

However, it is possible to figure out with the population problem. We can split the state columns to 49 states columns.