Data analysis on USA Road traffic Accidents

Nassima BENAMMAR

Context

- A road accident is related to several factors
  - weather
  - place
  - time
  - car
  - driver
  - ....
- Thank to Data science, the severity of road traffic accident can analysed and prevented
- Where does it occur the most?
- What feature has strong impact on the severity of the accident?
- Is driving by night dangerous? What about bad weather conditions?

Data source

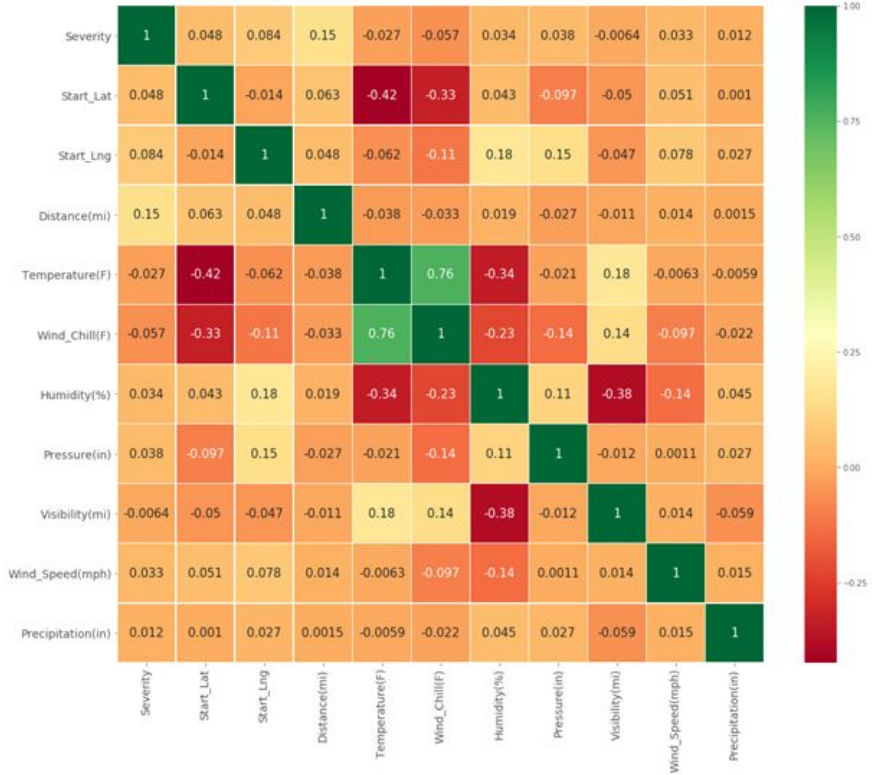
- We focus on the USA road accidents.
- Data source : https://smoosavi.org/datasets/us_accidents
- countrywide traffic accident dataset over 49 states of the USA from 2016 until June 2020
- 3513617 records
- 49 columns
- Data fields:
  - Description : ID, Source, time, coordinates, description, severity
  - Address : street, city, State, etc.
  - Weather : condition, temperature, visibility, humidity, wind speed
  - Traffic annotations : traffic signals, stop, junctions, crossing, etc.

Cleaning data

- Drop columns :
  - Attributes with a number of missing values much greater that the values.
  - Attributes that are not relevant for the prevention of accident such as nautical_twilight which gives the period of the day according to the nautic twilight.
  - Repetitive or unnecessary attributes such as the country
  - Low correlated attributes with the Severity
- Replace missing values
- Normalization
- 38 final columns

Correlation between quantitative columns

- No attribute has enough correlation with the Severity
- It is hard to predict using those attributes.

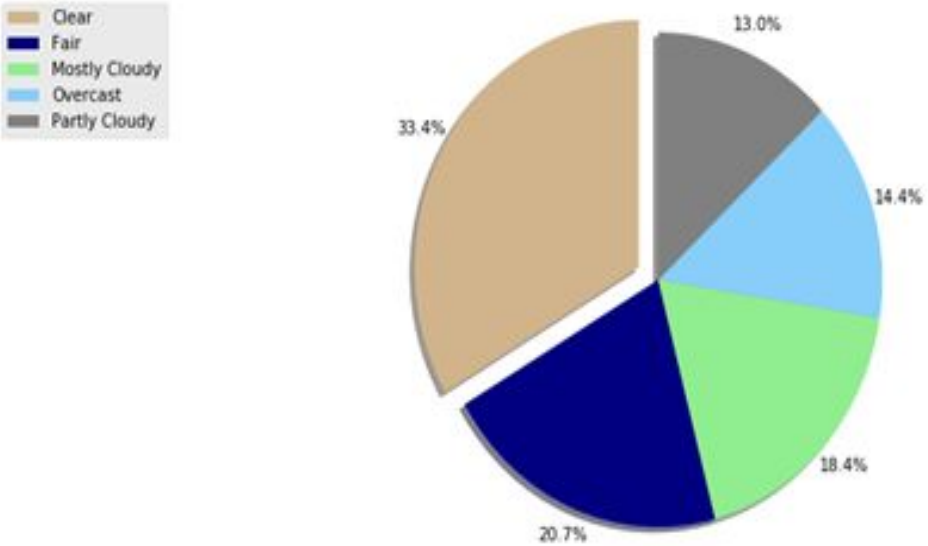


Data analysis by category of features

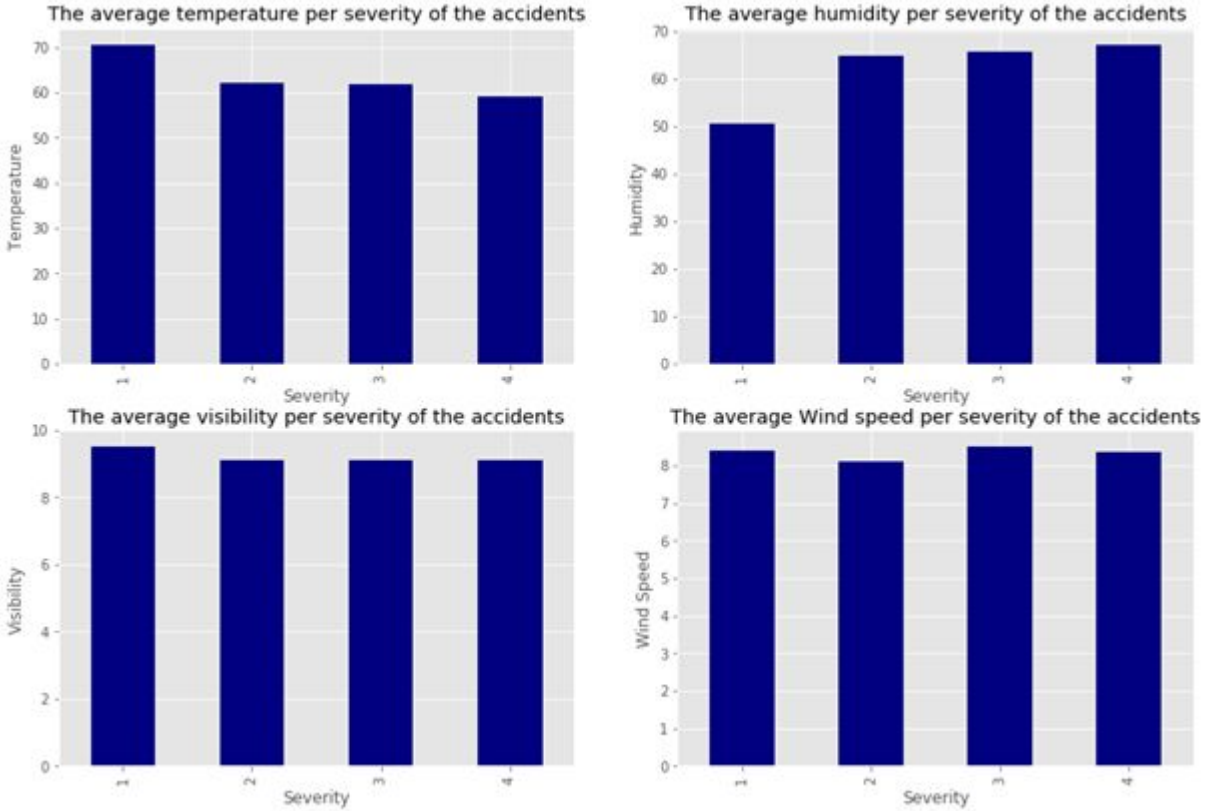
- Three categories
  - According to the weather
  - According to the time
  - According to the place

Data analysis on weather conditions

Number of accidents according to the weather condition



- Most of the accident occurs in a good weather condition (33,5% in clear weather and 20.5% in fair weather)



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

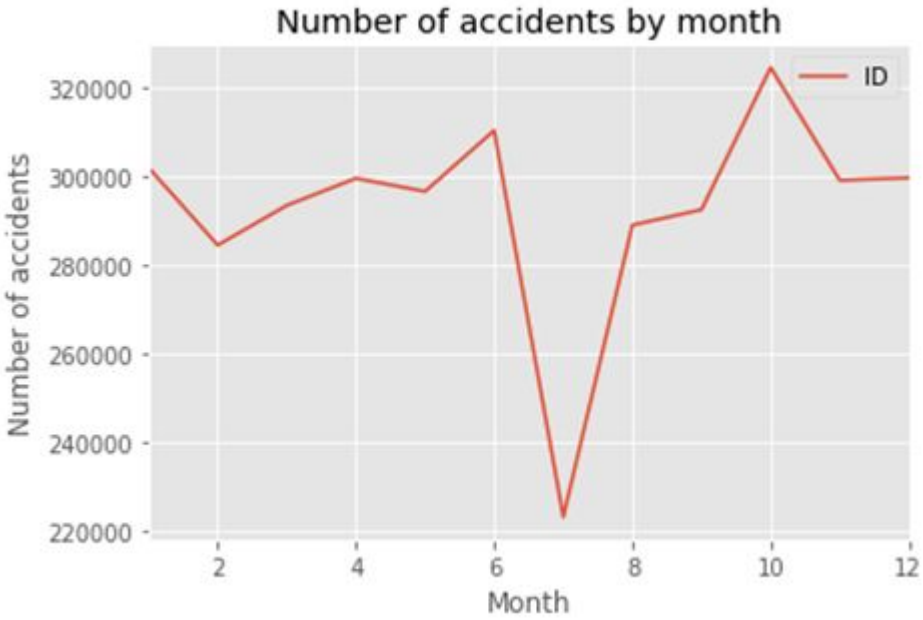
Data analysis on time

From 2016 to 2020

- The number of accidents increases from 2016 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**

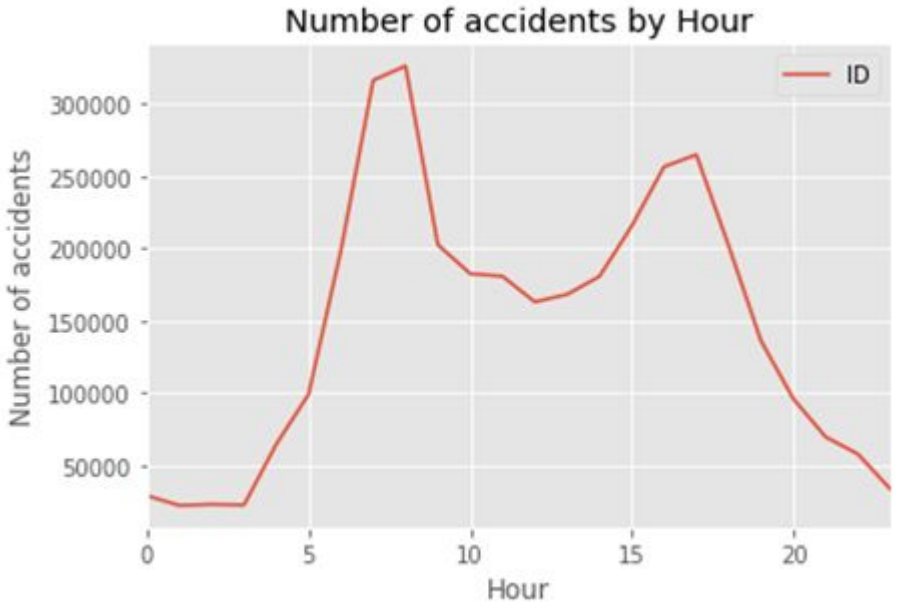
Number of accidents by month

- the greatest number of accidents occurs around October before
- it decreases in holiday period :
  - in Christmas holidays.
  - in summer holidays.
- Most of the accidents occur when people are actif



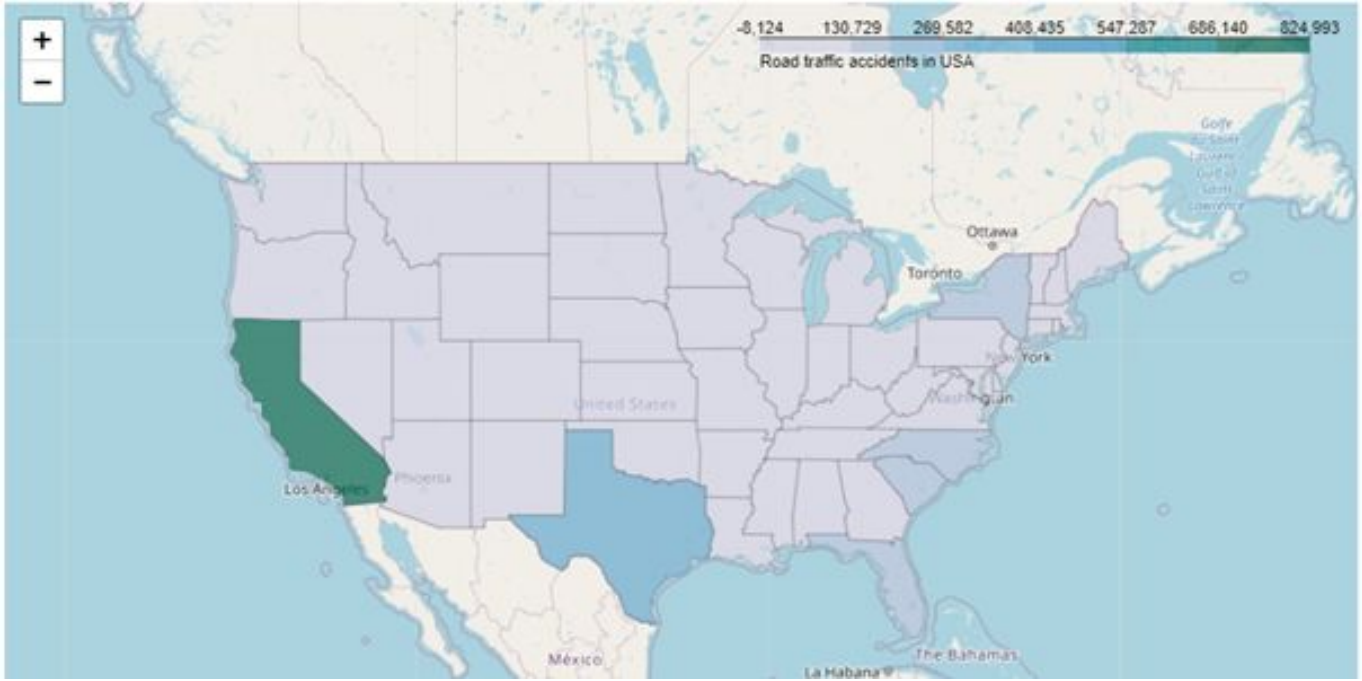
Number of accidents in different period of the day

- As for bad weather, people drive more carefully in bad condition : they drive carefully by night.



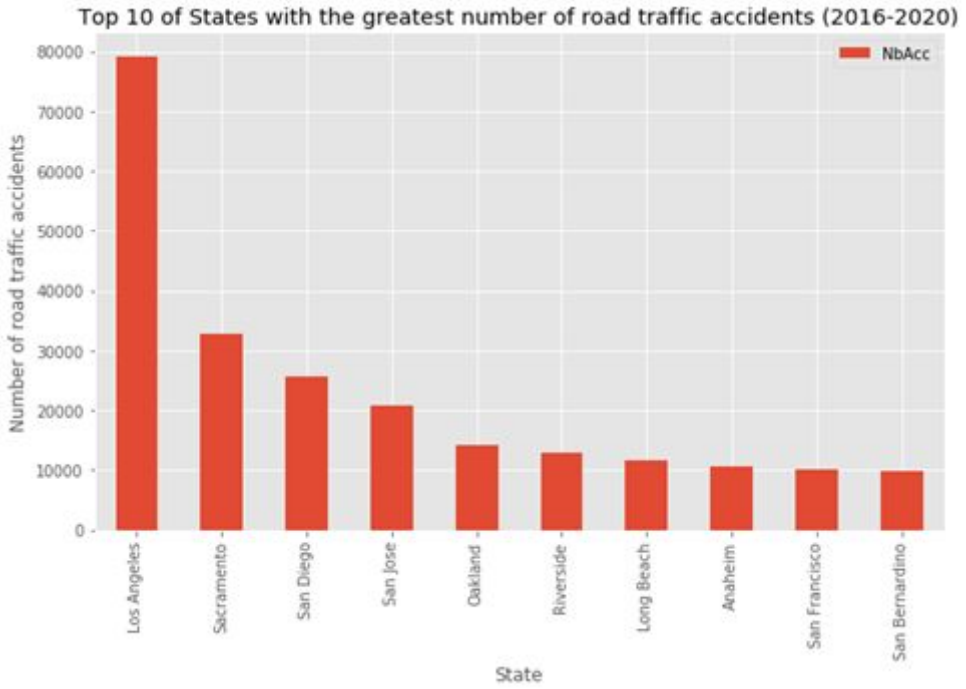
Data analysis based on the accident point

The states the most impacted by the road accidents



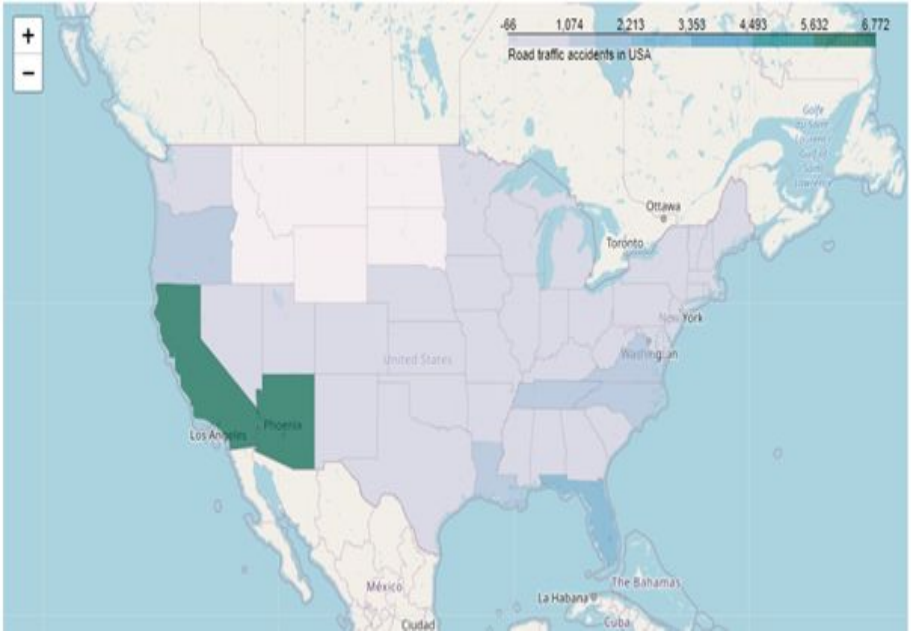
Focus on California

- Los Angeles is the cluster of danger in the road traffic accident.
- Sacramento and san diego in top 3.
- Cities around LA : oakland and Long Beach.
- San Francisco is in top 10 but much lower than other big cities of Calofornia.



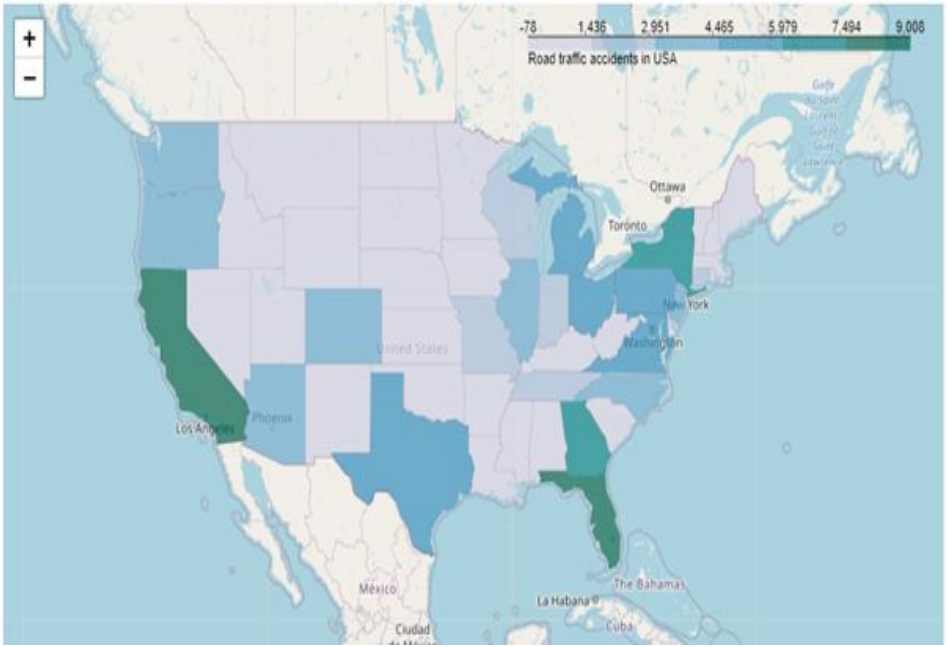
States with the greatest number of accidents with low severity

- Similar to the case of the number of total accidents by state.
- Arizona is in TOP 3.



States with the greatest number of accidents with high severity

- Severe accidents occur in the states where the population is dense.
- Focus on west coast and east coast
- Focus on Texas,
- Focus on Colorado because of Las Vegas city, Washington because of Seattle and Illinois because of Chicago.



Conclusion and future work

- Most of the accidents occur **in good weather condition, by daylight, in crowded places**.
- The severity of the accident is correlated to the **density of the population**.
- 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 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)
- Solution : split the state columns to 49 states columns.