US Accidents Project Proposal

Question/need:

The project will answer several questions about US car accidents:

- What is the number of accidents and their average severity during the past years?
- What times with the most accidents?
- What are the factors affecting the severity of the accident?
- What are the top state or county with the most accidents?
- Predict the number of accidents in NY at 2020 according to the visibility.
- Predict the number of accidents in 2021.
- Is there a problem in a specific street causing a lot of accidents?

Answering these questions will help both the police station and the municipality of the county to spread awareness and investigate the issues that causes the car accidents.

Data Description:

The dataset was obtained from Kaggle, this dataset has been collected in real-time it is about "US accidents", which covers 49 states of the USA. The dataset has about 1.5 million accident records and 47 features that contains information about the severity, start and end time of the accident, also had a more detailed about accident coordinate location, the length of the road extent affected by the accident, whether condition, visibility, humidity, wind direction and speed, shows the period of day (i.e. day or night) and the location (State, County, city, street name and number) from February 2016 to Dec 2020.

Short descriptions of each important features that will help to answer the questions:

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

Start_Time
Shows start time of the accident in local time zone.

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.

Start_Lat
Shows latitude in GPS coordinate of the start point.

Start_Lng
Shows longitude in GPS coordinate of the start point.

End_Lat
Shows latitude in GPS coordinate of the end point.

End_Lng
Shows longitude in GPS coordinate of the end point.

Distance(mi)
The length of the road extent affected by the accident.

Description
Shows natural language description of the accident

Number
Shows the street number in address record.

Street
Shows the street name in address record.

Side
Shows the relative side of the street (Right/Left) in address record.

City
Shows the city in address record.

County
Shows the county in address record.

State
Shows the state in address record.

County
Shows the county in address record.

Visibility(mi)
Shows visibility (in miles).

Weather_Condition
Shows the time-stamp of weather observation record (in local time).

Sunrise_Sunset
Shows the period of day (i.e. day or night) based on sunrise/sunset

If modeling, what will you predict as your target?

1-Predict the number of accidents in NY at 2020 according to the visibility.
2- Predict a number of expected accidents in 2021
3- Highlight the locations that it has high number of accidents **(Generative Model)**

Tools:

I will use this tools in the project:

- Data Processing: Pandas, Numpy.
- Modelling: SciKit-Learn, PyTorch.
- Visualization: Seaborn, Pygal.

MVP Goal:

I will submit with data cleaning and exploratory data analysis (EDA), in the EDA I will answer some of the listed questions of this proposal.