Road Accident Prediction for Drivers' Safety

Road accidents are main cause of injury and death globally. There are many factors that cause the road accident where weather condition is one of the important cause of the accident. Road accident can happen especially on bad weather or certain road features like bump or in junction. Weather condition like rain, fog can reduce the visibility and make the road slippery which lead to sever crashes and effect most of the people life especially during the peak hours of the days.

This project aim to find out how different weather condition, road features and time of the day effect the severity of the accident and smooth out the traffic flow . In this project I have use US Accident (2016-2023) data set which is available on Kaggle (https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data). The dataset contains 7.7 million accident records which covers 49 states of US (United State) from February 2016 to March 2023 (taken from part A report). For this project I have used the sampled data set which is provided in the Kaggle website which contain a 500,000 data set of the road accident with same structure, features and format of the original dataset. This project is available on Git hub (https://github.com/evashakya1/Big-Data-Analysis-and-Project.git)

Before our analysis I have cleaned the dataset to improve its quality. First I have checked whether the data set have any duplicate data and then ckecked for any missing value for the features. From Figure 1 we can see that End_Lat have about 44% of the missing data in the dataset.

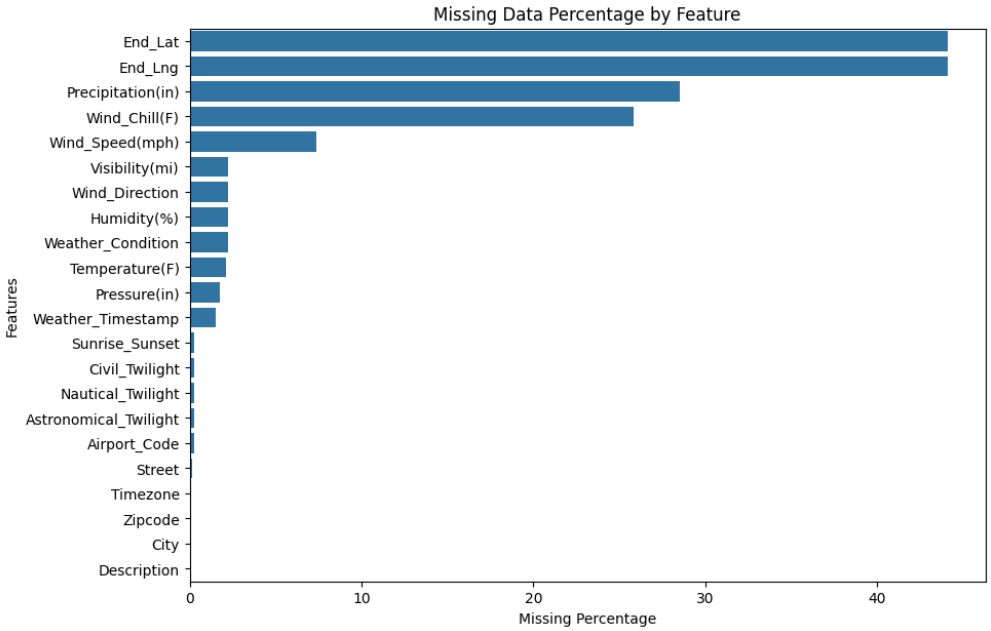

Figure 1. Missing Data of Features

But for this project, we will be working on only few features relevant to our topic. Therefore, if we remove all the missing value we might remove the important data as well so we first select the features that we will be working on this project. The features that we will be working on this project are Severity, Weather_Condition, Visibility(mi), Temperature(F), Humidity(%), Precipitation(in), Wind_Speed(mph), Wind_Direction, Start_Time, Sunrise_Sunset, State, City, Bump, Amenity, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal and Turning_Loop. After selecting the features, we notice that Precipitation(in) has the highest missing value about 28 % of missing value. When we remove all the missing Precipitation(in) values the number of missing values in other columns also drop. There was 500,000 data set before removing the Precipitation(in)'s missing data and 357,384 dataset after removing the missing value. From Figure 2, We can still see that there are several columns still contains missing values. So, I removed all the remaining missing values for better analysis.

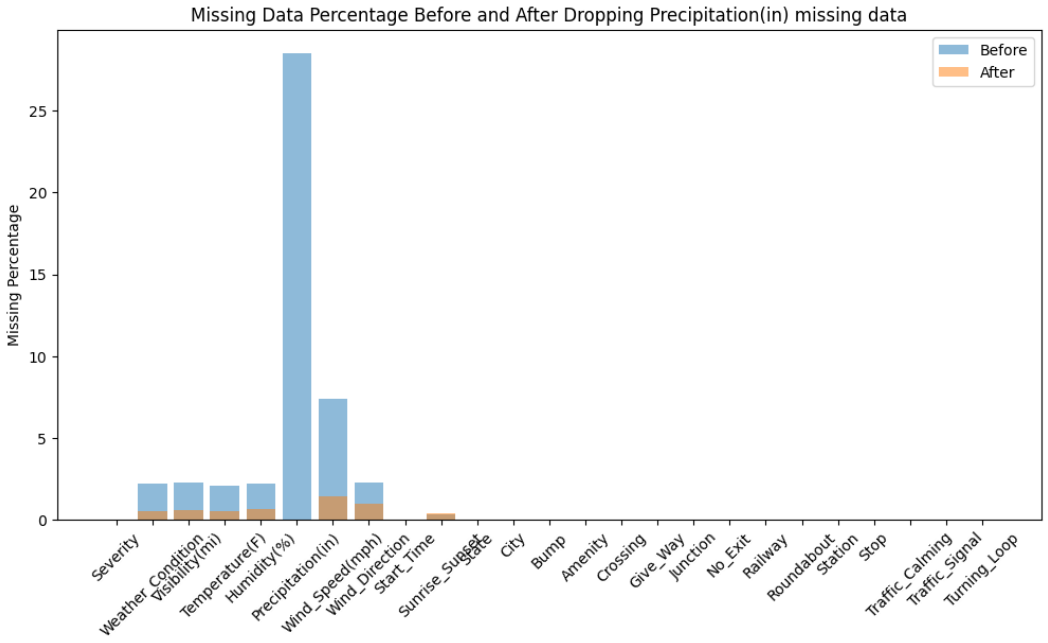

Figure 2. Missing data Before and After dropping Precipitation(in)

Now that our data is cleaned we can do some analysis. To support the exploratory data analysis(EDA) I convert the 'Start_Time' column to datetime format. This allows to extract the additional features from accident time such as hour of the day, day of the week, month and year for each records. These new time based features were added to the dataset to analyze the patterns in accident across the different time and days. From Bar Graph in Figure 3, we can clearly see that the number of accident increases steadily from 2016 reaching the peak in 2021 and slightly decreasing in 2022. The sharp drop in 2023 is more likely because the data for 2023 was incomplete.

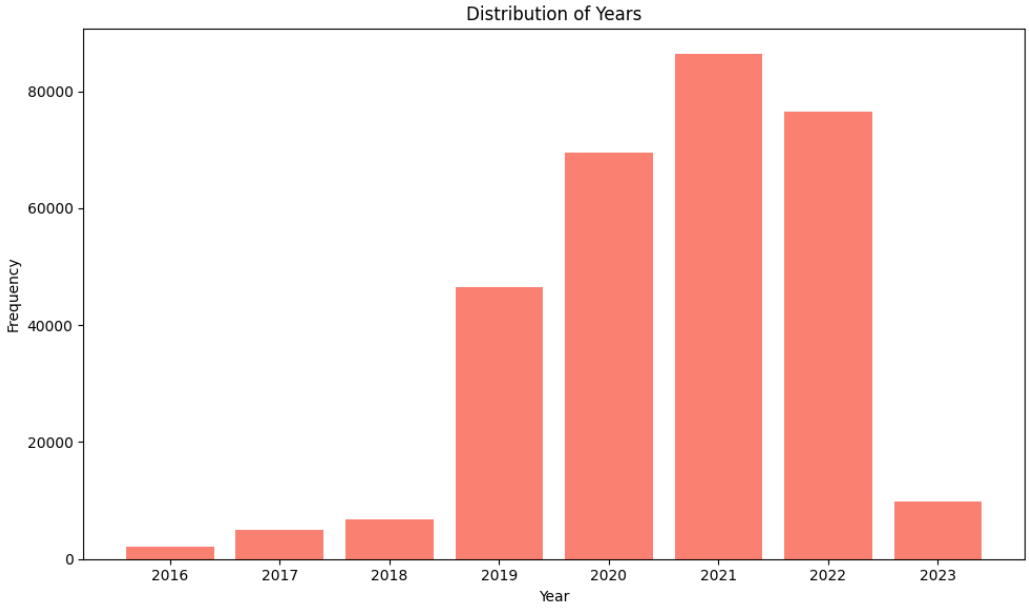

Figure 3. Distribution of Years

The Bar graph in Figure 4 shows the distribution of road accidents by hours of the day. The insight we get from this is the number of accident is low in the early morning till 5 am and increases at 6 am and with noticeeble peek around 7-8 am which indicated the morning rush hour. The accident again rise in the afternoon, reaching another peek between 4-6 pm which indicated the evening rush hour. This shows accidents are more likely to happen in high traffic especially during morning and evening.

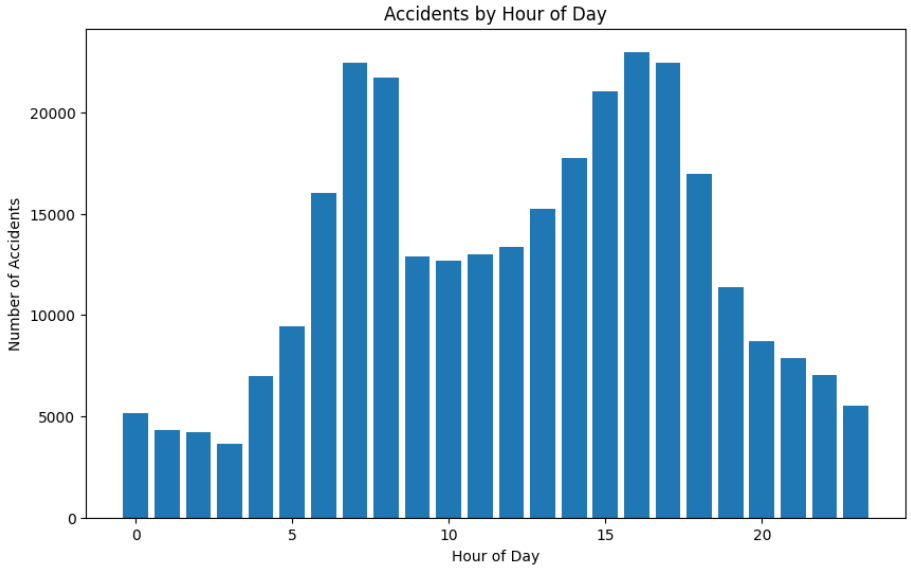

Figure 4. Accident by hour of Day

I would also like to know about what days the accident occurs most so that we can look into that particular days for the cause. The bar graph from the Figure 5, show the number of roads accidents

for each day of the week. We can notice that the accidents are most frequent on weekdays where Friday having the highest count. There is a noticeable drop on weekends where Saturday and Sunday have the lowest number of accidents. This shows that the accidents are more likely to happen in the working days might be due to high traffic volumes for work-relataed travels.

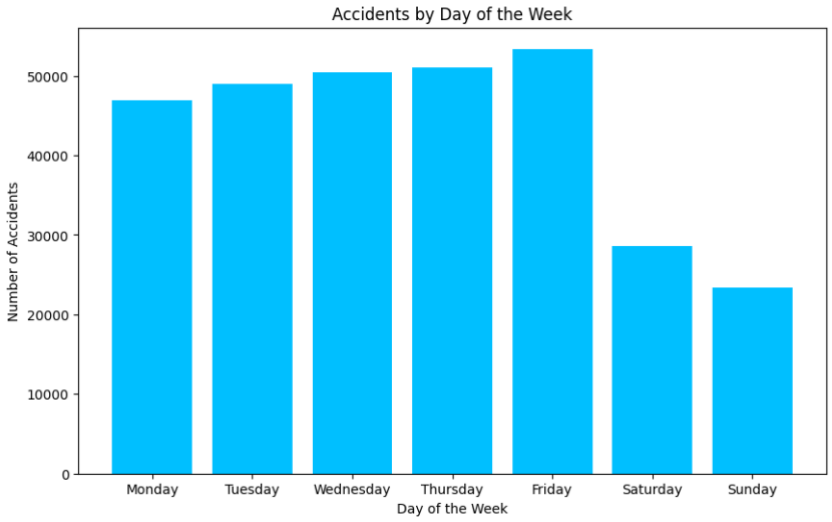

Figure 5. Accident by day of the week

Next, we look into the ratio of accident during day and night and what its severity under different weather conditions. Figure 6 shows the distribution of accident during day and night with its severity. It shows that the most accident occur during day. However at night there are still some of the accidents that have severity of 3 and 4 despite having few vehicle on the road. The “Severity” of each accident is recorded from 1 to 4 where 1 is minor accident with little impact on traffic and 4 being the most sever and dangerous accident which often causing major delay.

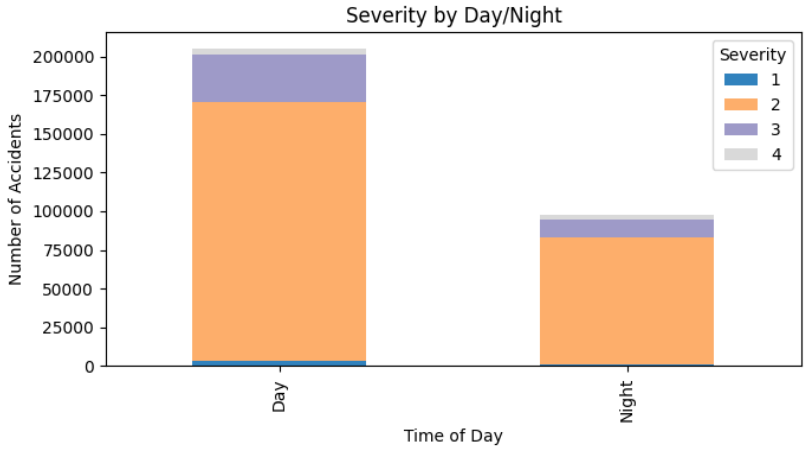

Figure 6. Severity by Day and Night

To further visualize about the different weather condition during the part of day we further explore the relationship between accident severity, weather condition and time of the day using heatmap. From

Figure 7, we can see the average severity of accidents increases significantly under different weather condition such as light rain and light snow, especially at night. Accident during the fair weather is less sever regardless of the time of the day. Figure 6 and 7, highlights that the major accident happen during the day and the risk of severe accident is higher at night especially during the poor weather like light rain and light snow where there is less visibility. Both adverse weather and low visibility plays a critical role in increasing road accident which indicates the extra need of caution while driving at night or in bad weather .

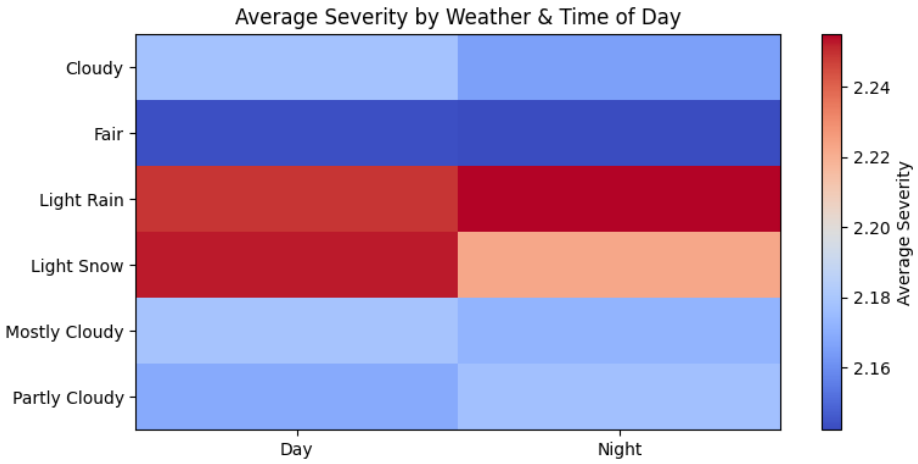

Figure 7. Average Severity by Weather and time of the day

In Figure 3, we notice that the number of accident is highest in the yeat 2021 compared to 2016. But what about the severity of the accident. If we look at the sever accident where 'Sever' refers to the accident with severity rate of 3 or 4 where it cause more delay due to some serious injuries which is important most important in road safety. From Figure 8 we can see that the decline in the percentage of severe accident over time. This suggest that while minor accidents have become more frequently reported, but sever accident have become less common over year. This could be due to improvement of the vehicular safety and some road infrastructure or can be due to the COVID-19 pandemic during which not only in US but globally the road accident has decrease significanly (Yasin, Grivna & Abu-Zidan 2021)

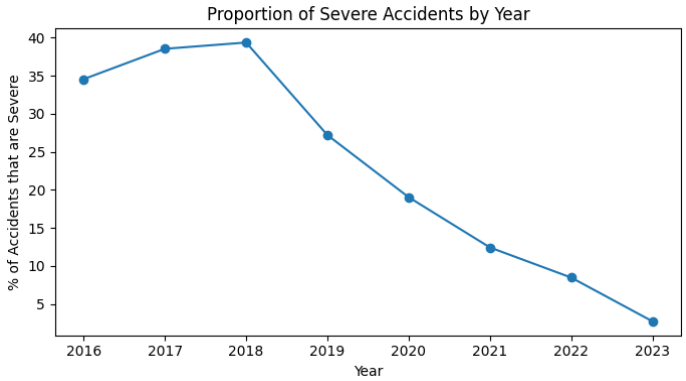

Figure 8. Severe accident by year

The pie chart in Figure 9 shows the overall distribution of the weather condition at the time of accident. 45% of the accident happened in Fair weather and followed by Cloudy, Most Cloudy and Party Cloudy. Other weather condition like rain , fog or light snow has the small proportion of the total accident. To further analysis according to the severity of the accident we can look into Figure 10.

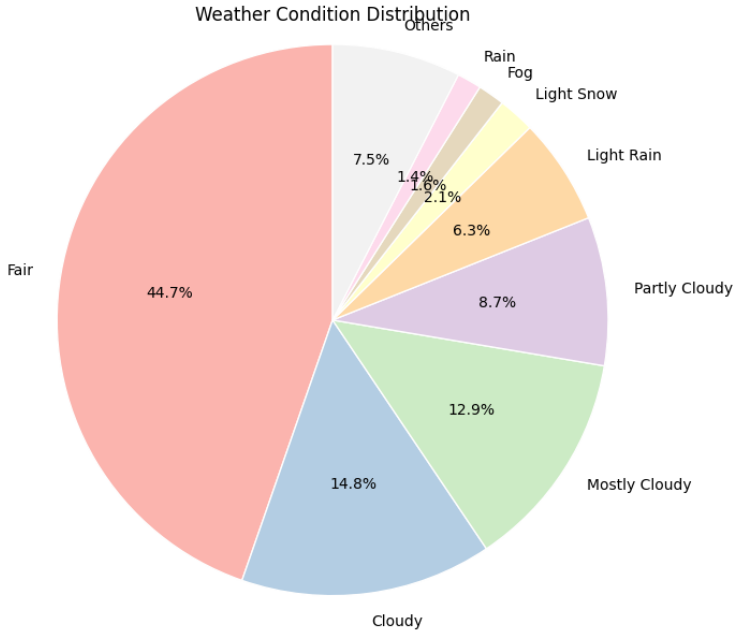

Figure 9. Weather Condition Distribution

Here in Figure 10 we can see the number of accident with both weather condition and severity level. Here we can see that the Fair weather has the highest number of accident where most of the accident are less severe(severity 2). But if we look closely on severity of accident (3 or 4 severity) it has the different story.

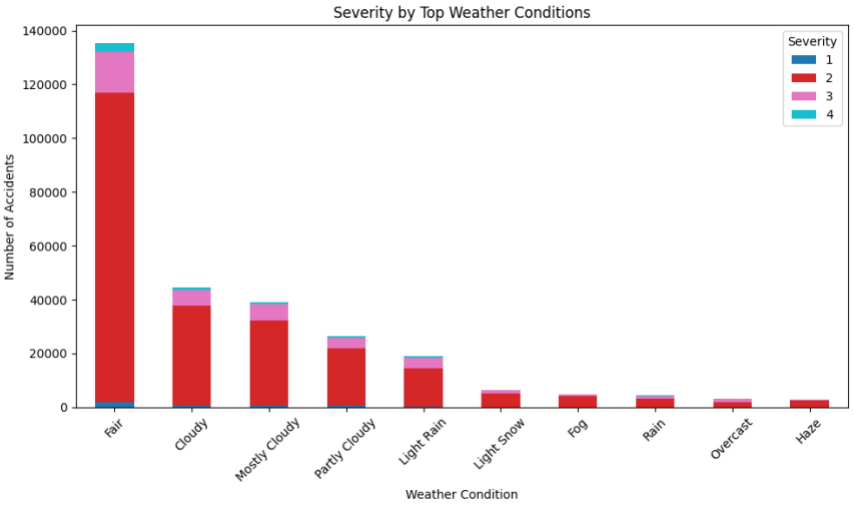

Figure10. Severity by Top Weather Condition

The bar chart below in Figure 11 shows the percentage of accident that are severe (severity 3 or 4). This shows that while the total number of the accident is high during fair weather, the proportion of severe accident is more in other condition such as 35% in Overcast, 25% in rain, 23% in light rain and 21% in light snow. Only 13% of the severe accident happened during fair weather. This shows that accident can happen in any weather but drivers are at high risk of having severe accident in weather like overcast, rain and snow.

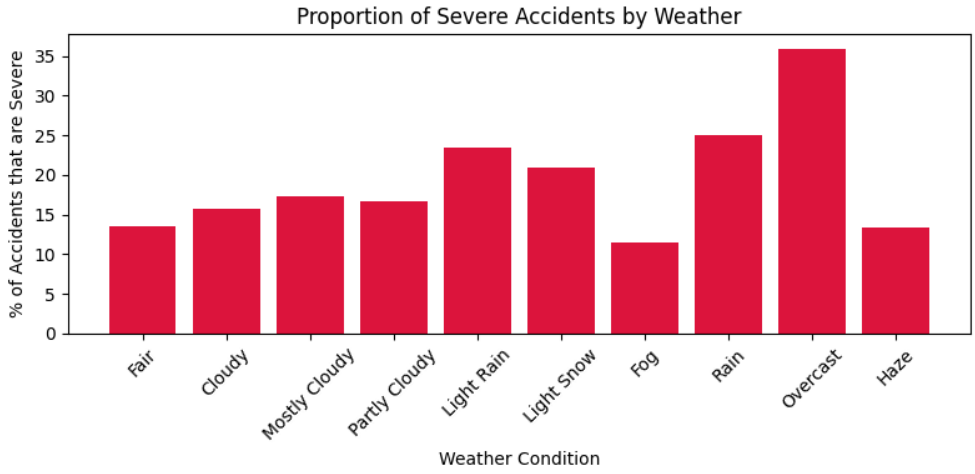

Figure 11. Sever Accident by Weather

Visibility is one of the important factor that can cause accident. From Figure 12, we can notice that the median of visibility is the same for all the different severity. A small number of sever accident are observe when the visibility is poor while indicate low visibility can be dangerous and also show that some severe accidents can also happen under good visibility condition. This suggest that visibility alone can not be consider a strong predictor of severity of accident but it should be considered along side other factor such as weather type and road condition.

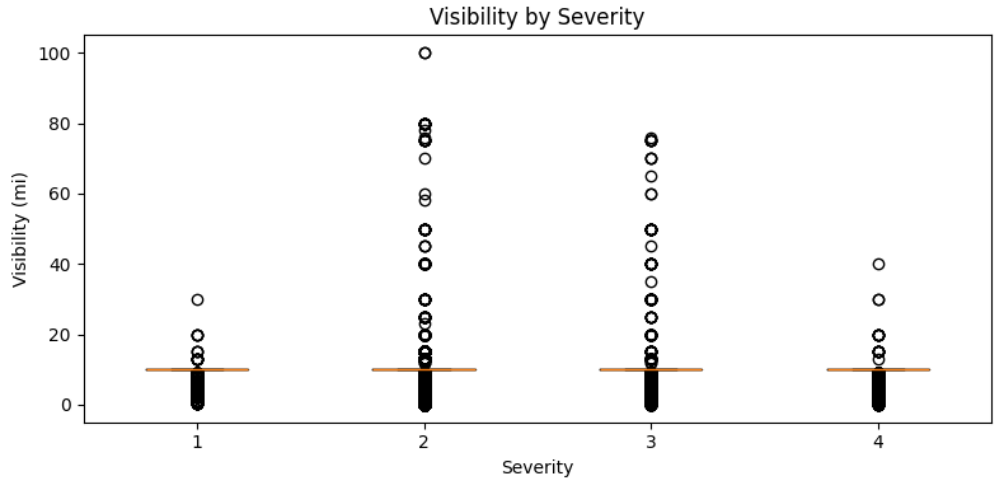

Figure 12. Visibility by Severity

Figure 13, shows the different relation between the different precipitation levels and accident severity. Most accident can occur when there is little or no precipitation regardless of the severity. From the Figure 13 the major data point are clustered around zero precipitation across all the severity which show that the heavy rain fall. Is very rare and may not be a dominant factor in most accident. However a few sever accidents (3 or 4 severity) do occur during the high precipitation as shown by the outliers in the Figure 13. This shows that while accidents are more common in dry condition the risk of sever accident may increase during heavy rainfall.

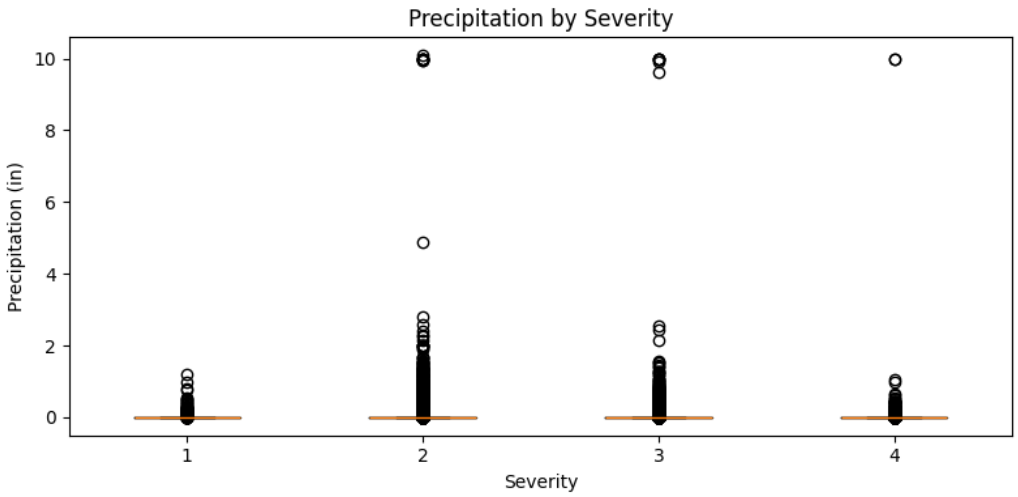

Figure 13. Precipitation by Severity.

For further analysis of the accident data I also anayze the various features of road on accident severity. The bar chart in the Figure 14 shows that the severe accident (severity 3 or 4) in different road features. Here we can see that certain features like Junction, Give_Way and Traffic_Calming devices are likely to have high proportion of sever accident. This suggest that beside weather condition specific road environment also plays a significant role in occurrence of the serious crashes. These patterns help to identify the high risk location and improve the road safety.

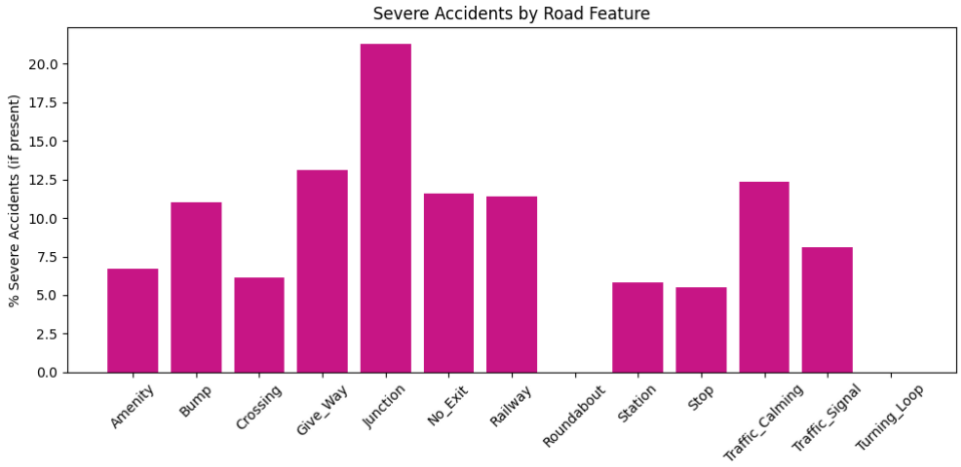

Figure 14. Severe Accident by Road Feature

To investigate the relationship between weather condition and accident severity feature like Weather_Condition, Visibility, time of day and road features as the input variable and Severity as the output. As severity is an ordered categorical variable this is a type of classification problem . The above analysis indicates that the weather and environment variable are the strong predictors for the classification of severity of accident. The bar graph, plots, pie chart and heatmaps are used to show the severity varies across different weather condition, road features and time period since the dataset is categorical and time base. To summarise, Weather related analysis reveled that the severe accidents are more likely to happen during rain or overcast while most of the accidents overall occur in fair weather. Similarly time based analysis highlight that during the rush hours and weekday the overall severity is high. Grouping accidents by road features showed that location like junction and traffic signals have higher proportion of sever accidents. These patterns shows that accident severity is not random but trends to cluster around weather condition, time of the day and road features. So to refine my research question on “How likely can different weather condition be used to predict the severity of road accident for driver’s safety?”(From Part A report) to “How different weather condition, road features and time of the day effect the severity of the accident?” this refinement is more focused and will help to improve the safety and reduce the impact of dangerous accidents.

References

1. Moosavi, S 2016, *US Accidents (2016 - 2023)*, Kaggle.com, viewed 6 July 2025, <https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents/data>.

2. *colab.google* 2025, colab.google, viewed 6 July 2025, <https://colab.google/>.

3. Yasin, YJ, Grivna, M & Abu-Zidan, FM 2021, 'Global impact of COVID-19 pandemic on road traffic collisions', *World Journal of Emergency Surgery*, vol. 16, Springer Science and Business Media LLC, no. 1, viewed 6 July 2025, <https://link.springer.com/article/10.1186/s13017-021-00395-8>.