Research question & data sourcing

Research question

The research question I have selected is “How do different weather conditions and road types influence the severity of traffic accidents?”. The research question I found describes the relationship between different weather conditions and road types and their effect on the severity of traffic accidents. Road accidents are a primary concern all over the world. The accident's severity will depend on conditions like weather and road types. In this modern world, transportation plays a significant role in human daily activities. All over the world, the traffic density has been increasing because of this modern transportation. Everyone focuses on their needs and wants, so they are not concerned about their and others' lives when driving on the roads.

Relevance

By investigating the factors that influence the occurrence and severity of traffic accidents, we can identify key areas for intervention and develop strategies to enhance road safety. Understanding the impact of weather conditions, road types, and driver behaviour can help policymakers and traffic authorities implement targeted measures to reduce accident rates and improve overall traffic management. So, the research will guide us in analyzing this topic further.

Source

I found a dataset called ‘Traffic accident predictions’ from Kaggle,

https://www.kaggle.com/datasets/denkuznetz/traffic-accident-prediction

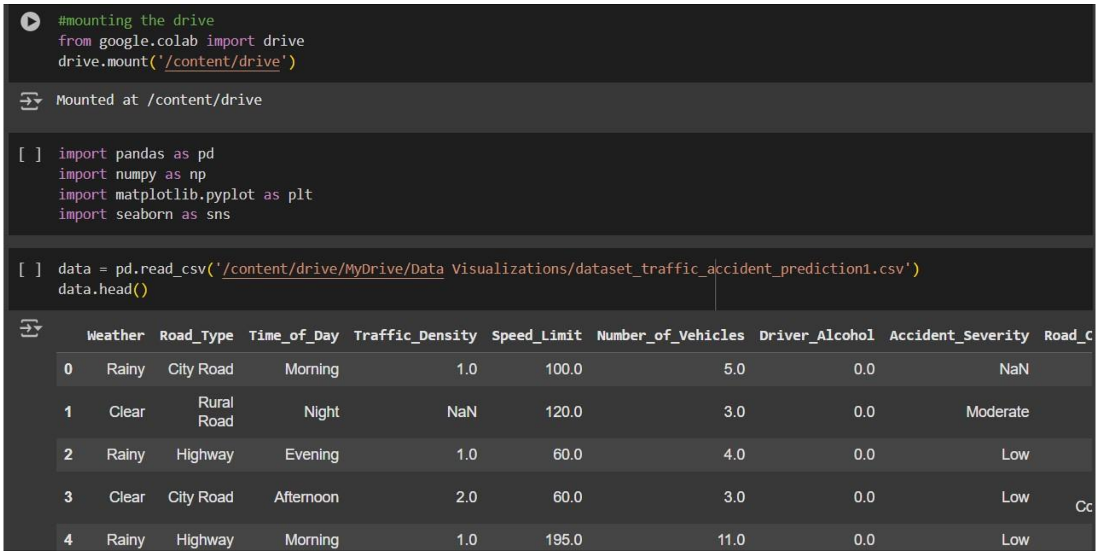
Data Preparation

Data preparation is the step that converts our raw dataset into a tidy format. Tidy format will evaluate the dataset's quality and improve our dataset's further insight.

Variables and observations

- Weather: The weather conditions at the time of the accident.
- Road_Type: The type of road where the accident occurred.
- Time_of_Day: The time of day when the accident happened.
- Traffic_Density: The traffic density on a scale from 0 to 2 at the time of the accident.
- Speed_Limit: The speed limit on the road where the accident occurred, measured in km/h.
- Number_of_Vehicles: The number of vehicles involved in the accident.

- Driver_Alcohol: Indicates whether the driver had consumed alcohol (0 for driver used alcohol, 1 for driver didn’t use alcohol)
- Accident_Severity: The severity of the accident (Low, Moderate, High) • Road_Condition: The condition of the road at the time of the accident.
- Driver_Age: The age of the driver involved in the accident.
- Driver_Experience: The driver's driving experience is measured in years.
- Road_Light_Condition: The lighting condition on the road at the time of the accident.
- Accident: Indicates whether an accident occurred. (0 for accident happen, 1 for accident not happen)



```
#mounting the drive
from google.colab import drive
drive.mount('/content/drive')
```

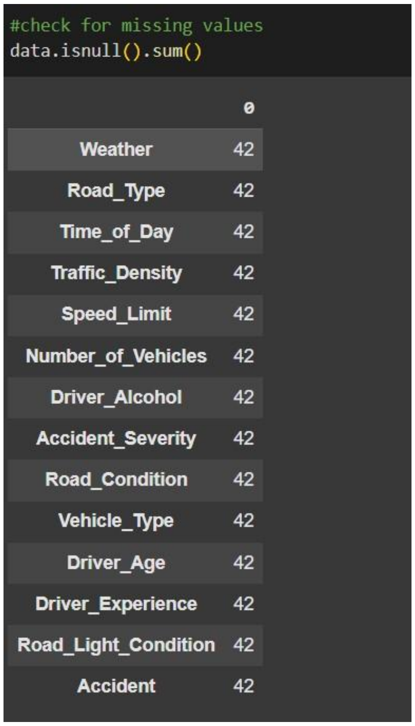
```
Mounted at /content/drive
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

```
data = pd.read_csv('/content/drive/MyDrive/Data Visualizations/dataset_traffic_accident_prediction1.csv')
data.head()
```

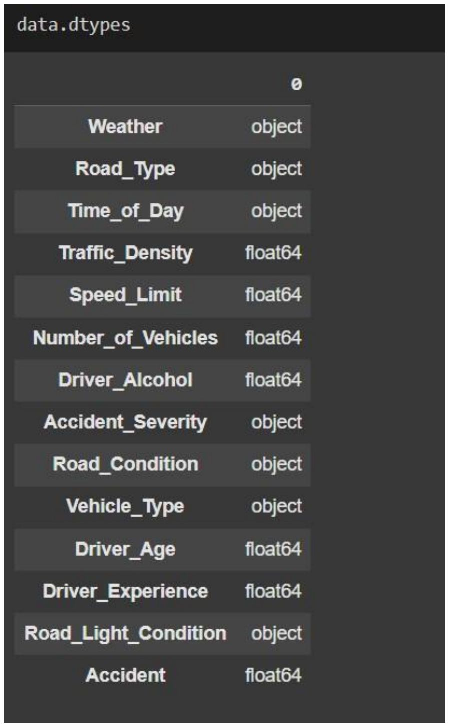
	Weather	Road_Type	Time_of_Day	Traffic_Density	Speed_Limit	Number_of_Vehicles	Driver_Alcohol	Accident_Severity	Road_C
0	Rainy	City Road	Morning	1.0	100.0	5.0	0.0	NaN	
1	Clear	Rural Road	Night	NaN	120.0	3.0	0.0	Moderate	
2	Rainy	Highway	Evening	1.0	60.0	4.0	0.0	Low	
3	Clear	City Road	Afternoon	2.0	60.0	3.0	0.0	Low	Cc
4	Rainy	Highway	Morning	1.0	195.0	11.0	0.0	Low	

The raw dataset was assigned as a ‘data’ variable. The raw dataset is already tidy, but there were missing values, and the data types of some variables were incorrect. So, I filled out the missing values and corrected those data types under the data preparation step.



```
#check for missing values
data.isnull().sum()
```

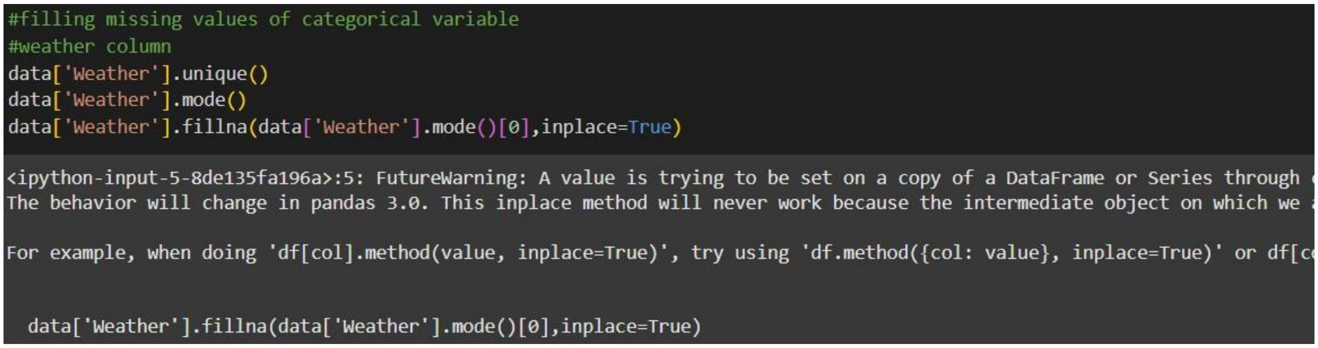
```
                     0
Weather             42
Road_Type           42
Time_of_Day         42
Traffic_Density     42
Speed_Limit         42
Number_of_Vehicles  42
Driver_Alcohol      42
Accident_Severity   42
Road_Condition      42
Vehicle_Type        42
Driver_Age          42
Driver_Experience   42
Road_Light_Condition 42
Accident            42
```



```
data.dtypes
```

```
                          0
Weather              object
Road_Type            object
Time_of_Day          object
Traffic_Density     float64
Speed_Limit         float64
Number_of_Vehicles  float64
Driver_Alcohol      float64
Accident_Severity    object
Road_Condition       object
Vehicle_Type         object
Driver_Age          float64
Driver_Experience   float64
Road_Light_Condition object
Accident            float64
```

The ‘is null ().sum()’ gave all the missing value counts for each variable. The ‘dtypes’ gave the data type of each variable.



```
#filling missing values of categorical variable
#weather column
data['Weather'].unique()
data['Weather'].mode()
data['Weather'].fillna(data['Weather'].mode()[0],inplace=True)
```

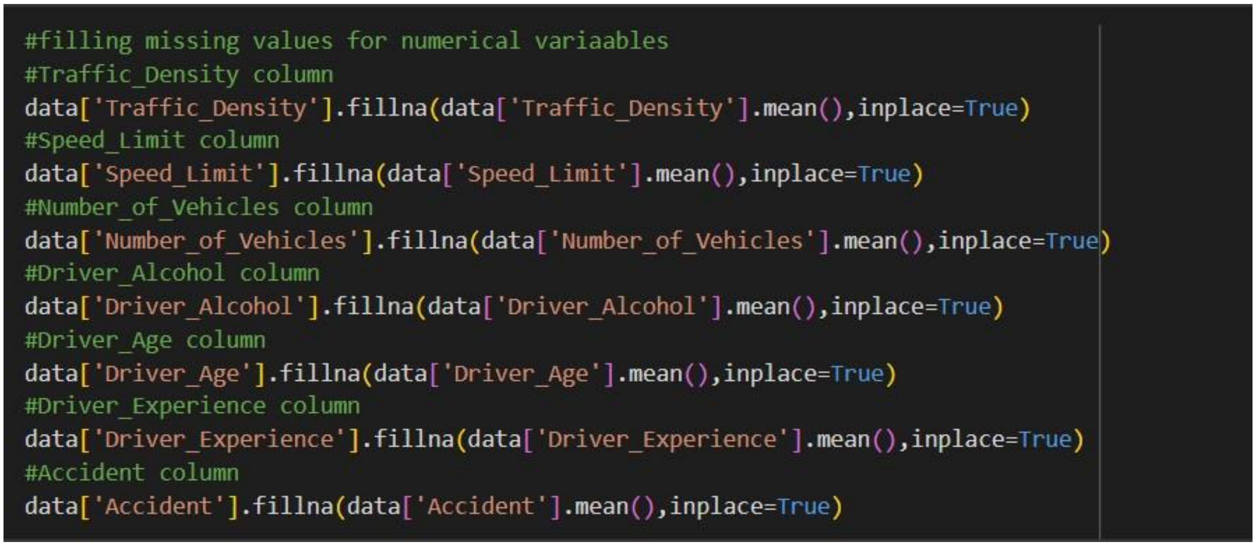
```
<ipython-input-5-8de135fa196a>:5: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through
The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[c


  data['Weather'].fillna(data['Weather'].mode()[0],inplace=True)
```

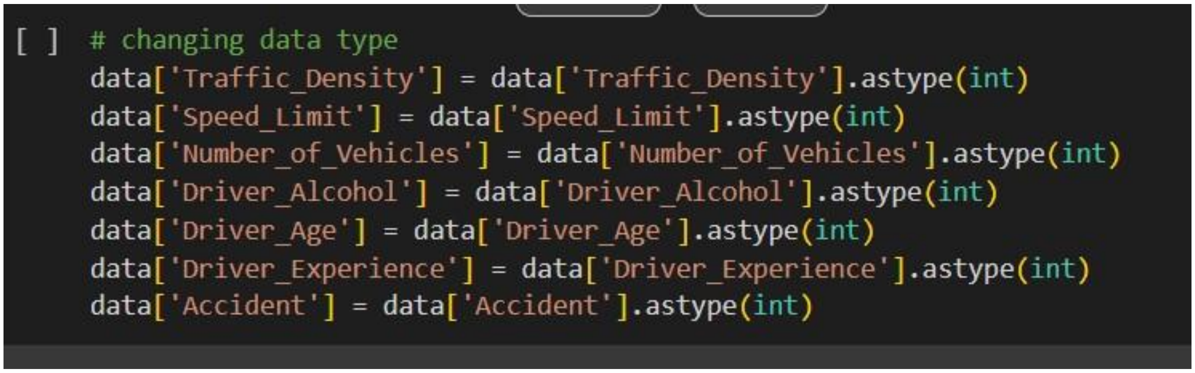
The ‘Weather’ variable is categorical, so the above code is executed using ‘mode()’ to fill in the missing value of the categorical variable. The ‘mode’ evaluates the most frequently used value in this variable, and according to this code, the missing values are filled using the ‘mode’.

For all categorical variables, the above code is executed.



```
#filling missing values for numerical variaables
#Traffic_Density column
data['Traffic_Density'].fillna(data['Traffic_Density'].mean(),inplace=True)
#Speed_Limit column
data['Speed_Limit'].fillna(data['Speed_Limit'].mean(),inplace=True)
#Number_of_Vehicles column
data['Number_of_Vehicles'].fillna(data['Number_of_Vehicles'].mean(),inplace=True)
#Driver_Alcohol column
data['Driver_Alcohol'].fillna(data['Driver_Alcohol'].mean(),inplace=True)
#Driver_Age column
data['Driver_Age'].fillna(data['Driver_Age'].mean(),inplace=True)
#Driver_Experience column
data['Driver_Experience'].fillna(data['Driver_Experience'].mean(),inplace=True)
#Accident column
data['Accident'].fillna(data['Accident'].mean(),inplace=True)
```

The above screenshot will explain all the missing values in the numerical variables. The 'mean()' function gave the average value of each variable, and this above code evaluates the missing value filled with the mean value.



```
[ ]  # changing data type
     data['Traffic_Density'] = data['Traffic_Density'].astype(int)
     data['Speed_Limit'] = data['Speed_Limit'].astype(int)
     data['Number_of_Vehicles'] = data['Number_of_Vehicles'].astype(int)
     data['Driver_Alcohol'] = data['Driver_Alcohol'].astype(int)
     data['Driver_Age'] = data['Driver_Age'].astype(int)
     data['Driver_Experience'] = data['Driver_Experience'].astype(int)
     data['Accident'] = data['Accident'].astype(int)
```

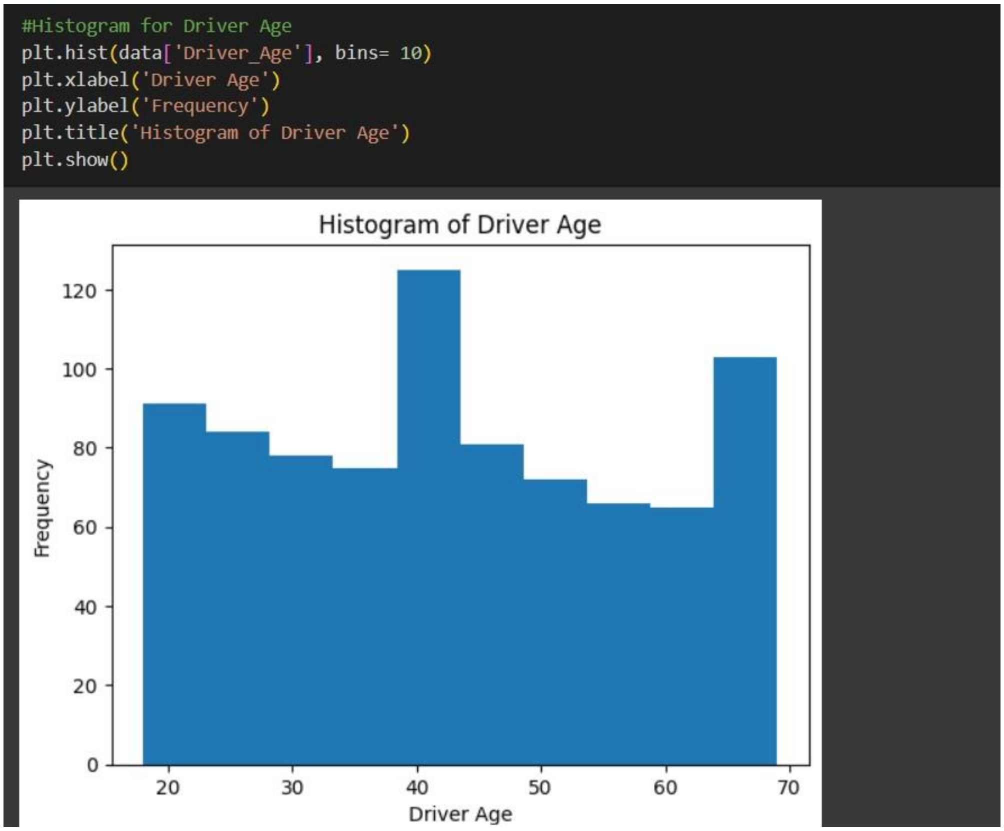
The other data preparation method is the above code, which evaluates the changing of data types. All numerical variables were 'float' type, and according to the intention of this code, the float data types were changed to the integer type.

Exploratory Data Analysis

This is a crucial step in the data analysis process. It involves summarizing the main features of the dataset, often using visual methods. EDA is used to understand the data's distribution, detect patterns and check assumptions with the help of summary statistics and graphical representations.

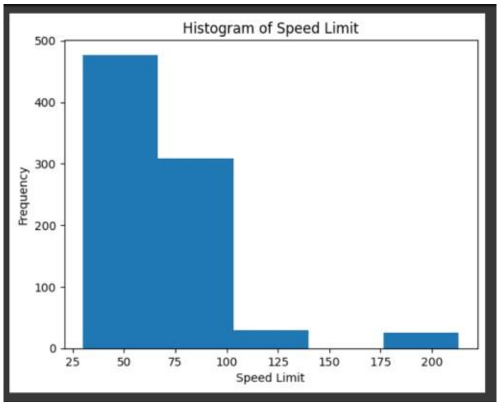
Univariate Analysis

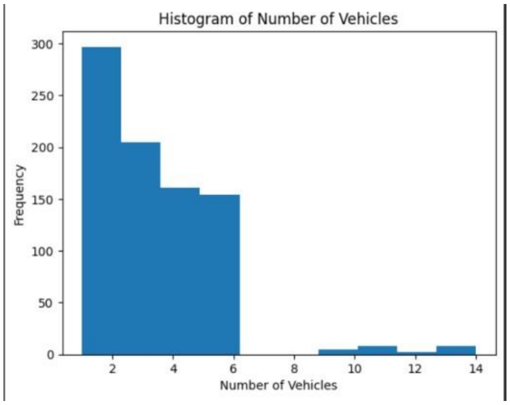
- **Distribution for Numerical variables**

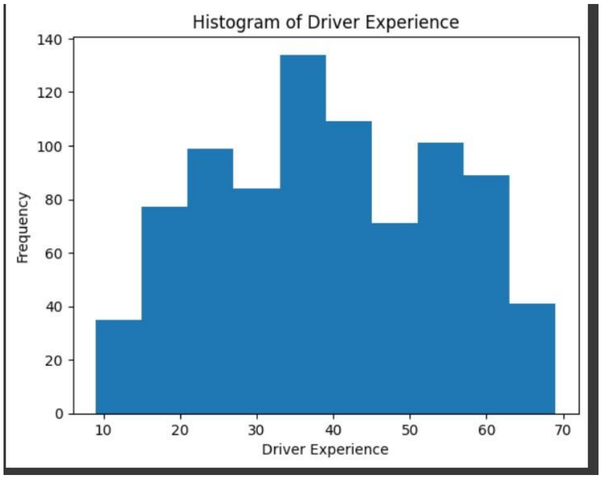


```
#Histogram for Driver Age
plt.hist(data['Driver_Age'], bins= 10)
plt.xlabel('Driver Age')
plt.ylabel('Frequency')
plt.title('Histogram of Driver Age')
plt.show()
```

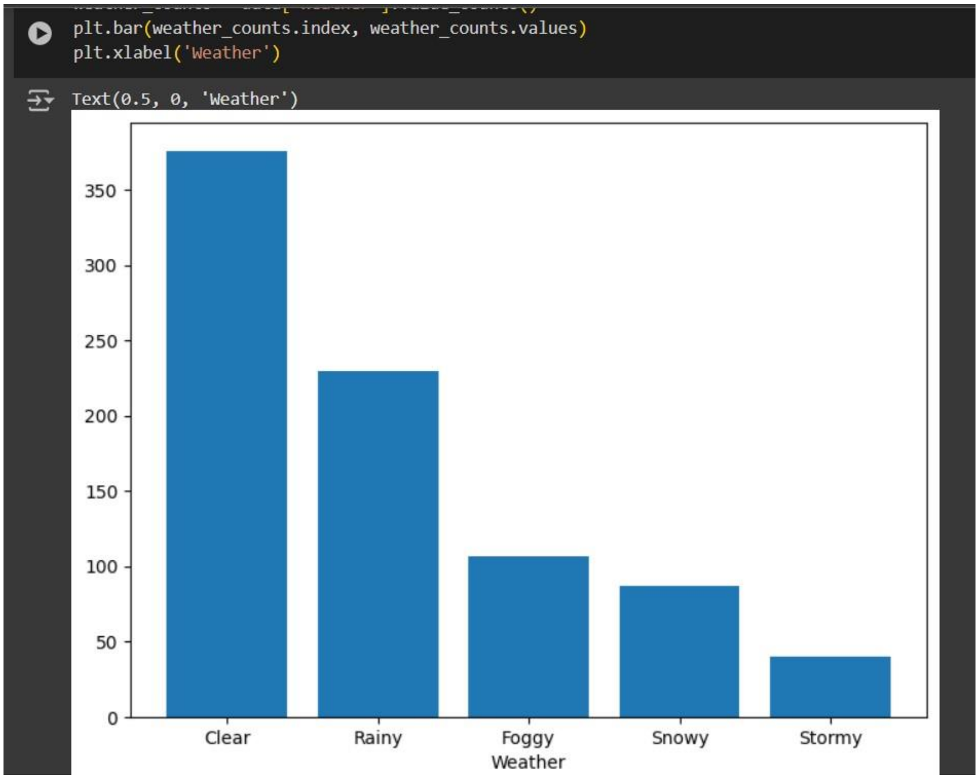
The histogram visualization is evaluated to check the distribution of numerical variables. The above histogram shows the distribution of the 'Driver_Age' variable. According to this distribution, the minimum age in this prediction dataset is 20, and the maximum is 70. The shape of this distribution looked like a normal distribution. For all other numerical variables, the above histogram is executed to check the distribution of each variable.






Histogram of Driver Experience
Frequency
Driver Experience
0
20
40
60
80
100
120
140
10
20
30
40
50
60
70

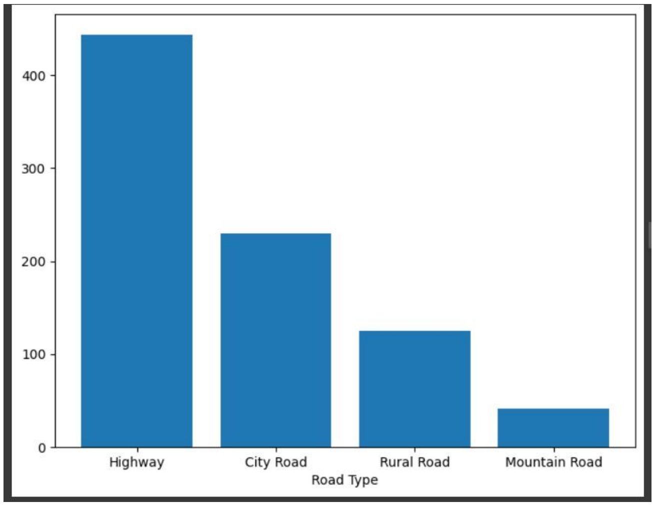
- **Distribution of Categorical variables**

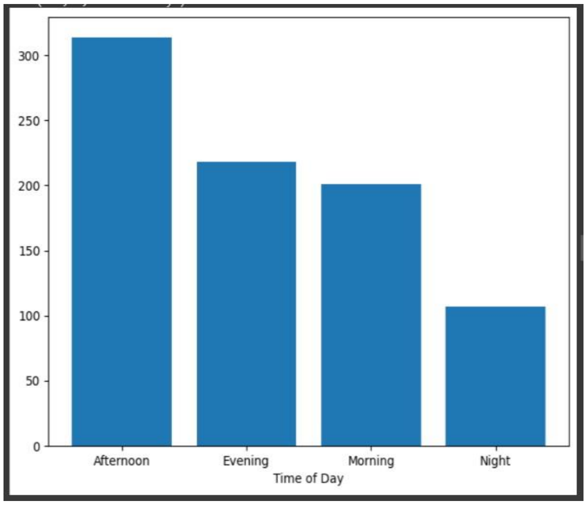


```
plt.bar(weather_counts.index, weather_counts.values)
plt.xlabel('Weather')
```

```
Text(0.5, 0, 'Weather')
```

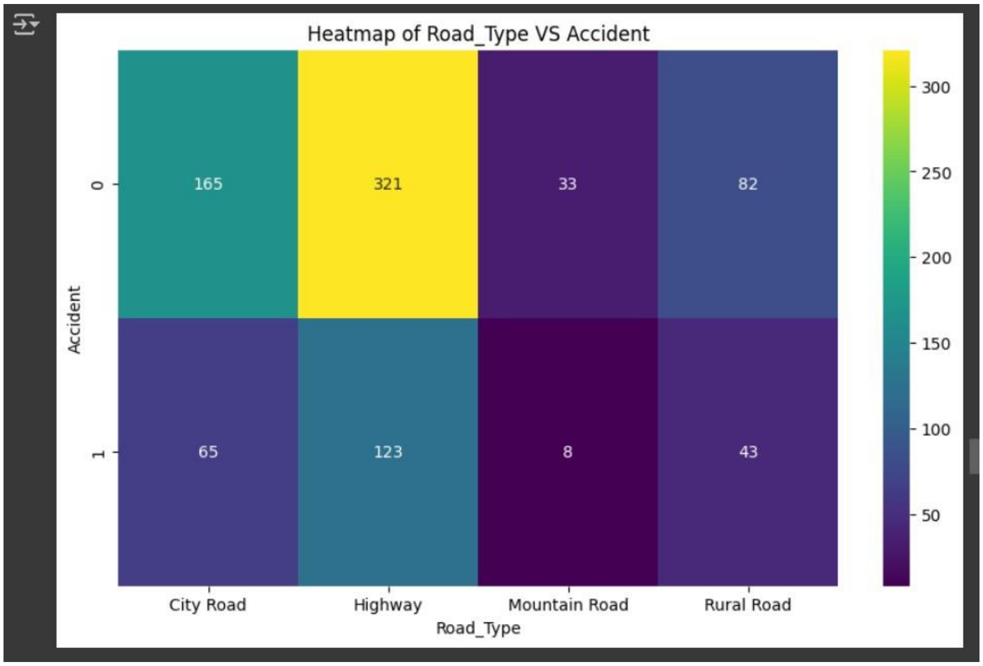
The bar graph is evaluated to check the distribution of the 'Weather' variable. This variable contains five weather conditions: Clear, rainy, foggy, snowy, and stormy. The bar graph shows the frequency of each weather condition in this variable. The same method is used for other categorical variables.



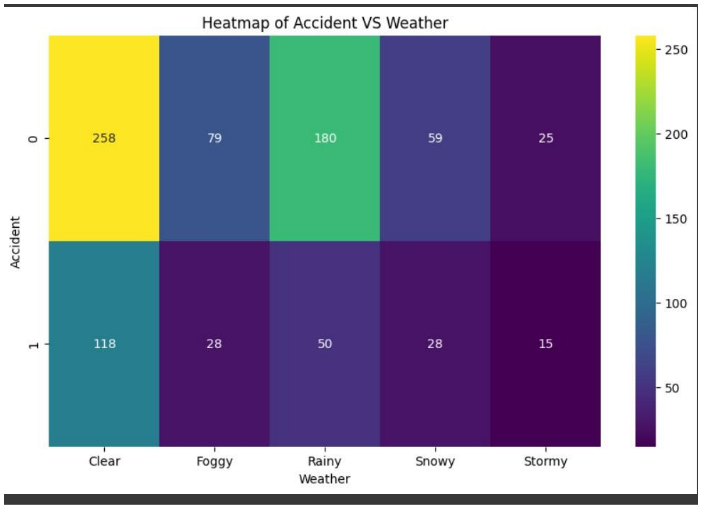


Bivariate Analysis

- **Heat Map**

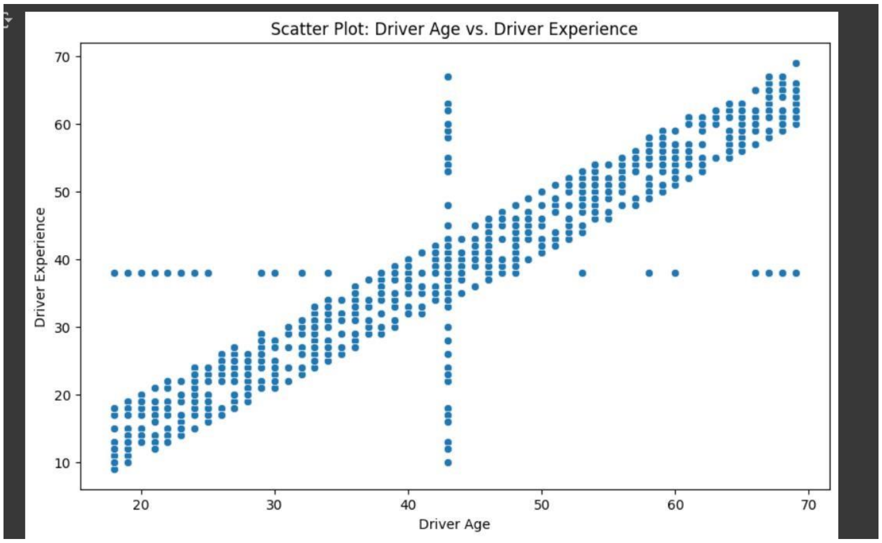


The above heat map indicates the probability of accidents by the type of road. In here, the '0', the accident will occur according to the predictions, and the '1' represents the accident will not occur according to the projections. These observations suggest that different road types, like city roads, highways, Mountain roads, and rural roads, may pose different risks to road users.



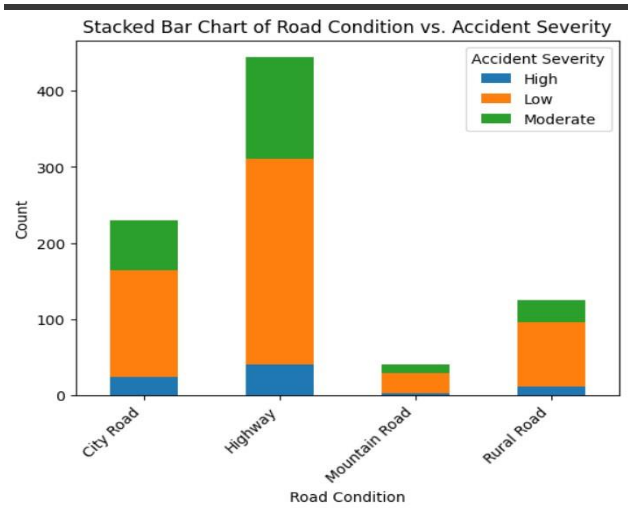
This observation suggests that weather conditions may pose different risks to road users.

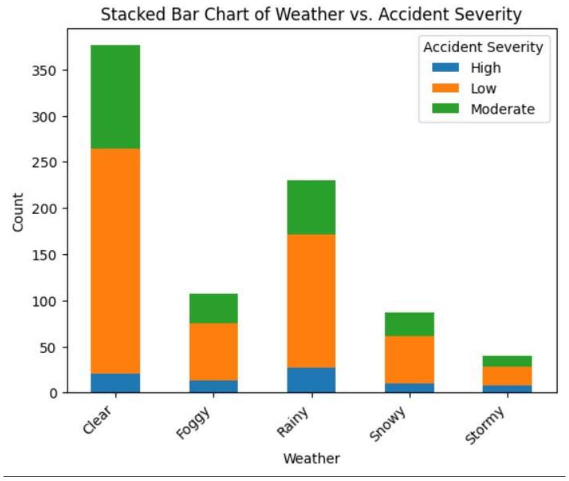
- **Scatter plot**



The above scatter plot indicates the relationship between the driver's age and experience. •

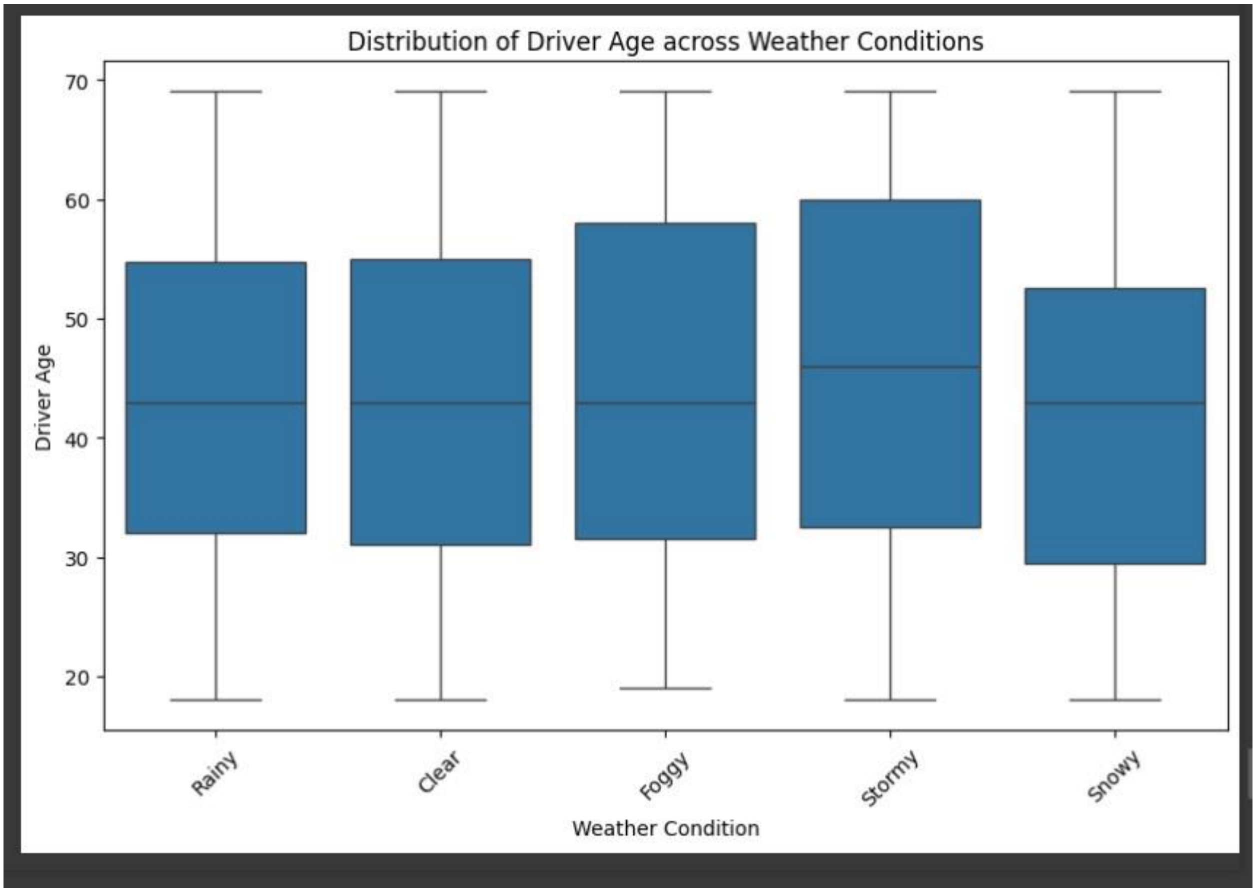
Staked bar graph



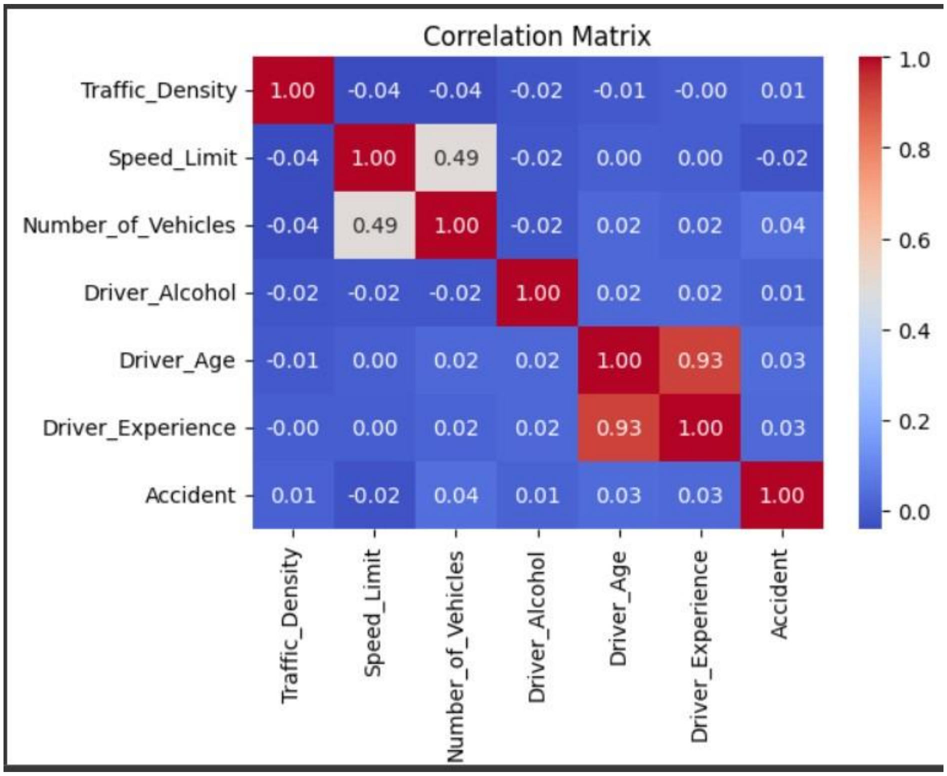


These two variables are the main targeted factors according to my research problem. The accident severity is a categorical variable, but here it's counted the records from each severity type. Stacked bar graphs are reveling patterns and correlations between the two variables. For instance, if a particular secondary category consistently occupies a larger proportion of a specific primary category's bar, it suggests a potential relationship between those categories. For example, in the Road Conditions vs accident severity graph, accident severity is higher in highway road types, and the severity of accidents is high in clear conditions.

- **Box plot**


Distribution of Driver Age across Weather Conditions
Driver Age
70
60
50
40
30
20
Rainy
Clear
Foggy
Stormy
Snowy
Weather Condition

- **Multivariate Analysis**

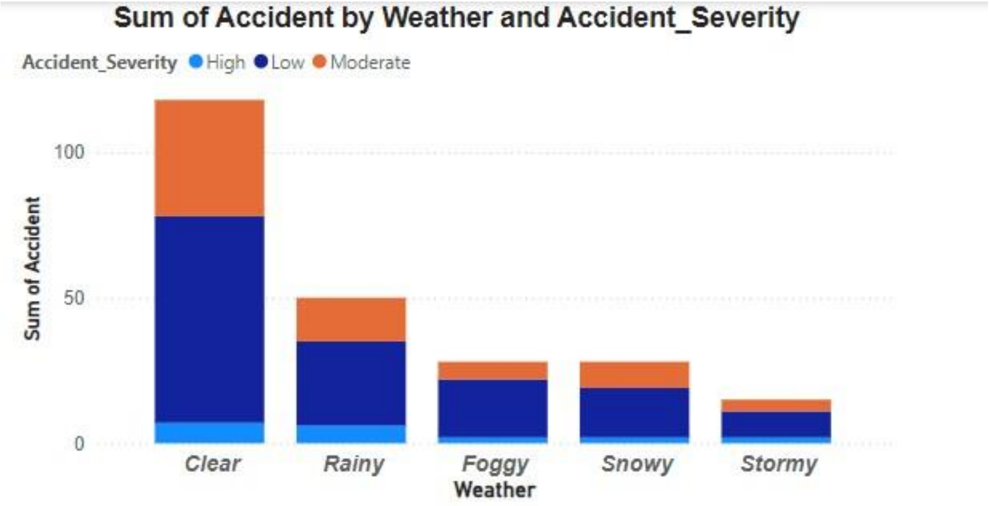


The correlation matrix is the best method for multivariate analysis. It helps to understand the relationships between the numerical variables, identifying which variables are firmly related, weekly related or unrelated.

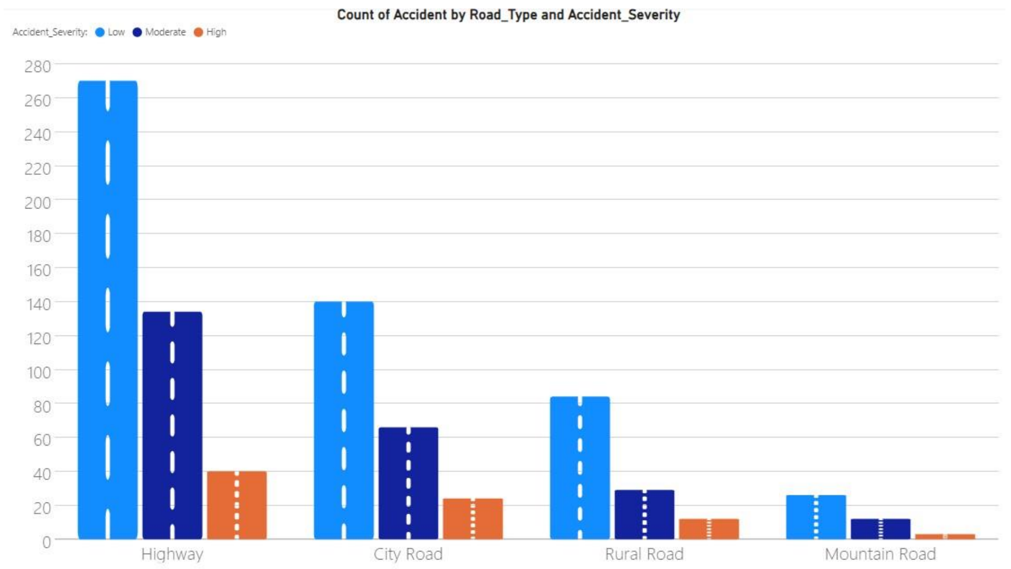
Data Storytelling

Every year, thousands of lives are lost on the roads, and traffic accidents impact countless more. But do these accidents happen without any reason, or are there any hidden patterns we can uncover to prevent them? Let's dive into the data to find out. The research question is How do different weather conditions and road types influence the severity of traffic accidents?". In response to this research question, the dataset I found will explain the factors contributing to traffic accidents. This dataset contains key details about the weather, road conditions, traffic density, and driver alcohol level.

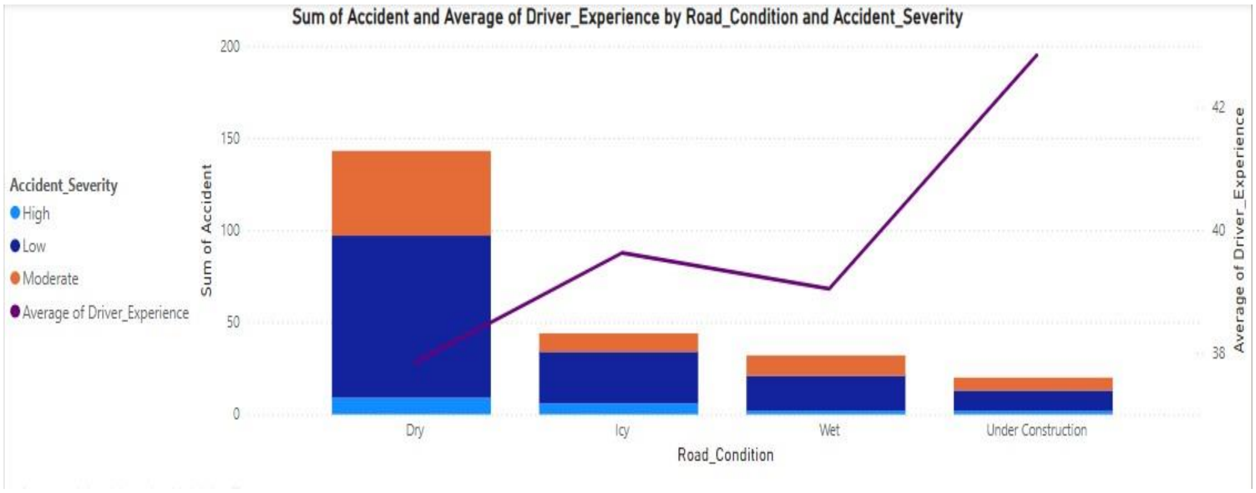
According to the EDA step, I realized that most of the accidents occurred, and the severity was high because of the road type, weather conditions, and alcohol level of the drivers.



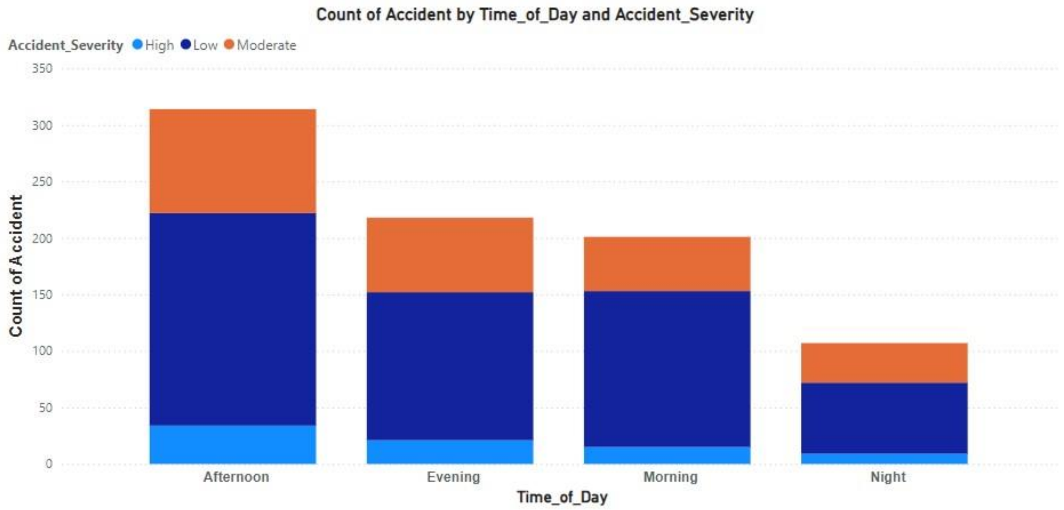
Do you think weather conditions don't impact accidents? The above graph will explain how weather conditions impact accidents. Surprisingly, clear weather conditions cause the highest number of traffic accidents. This might seem counterintuitive, but clear conditions lead to overconfidence, speeding and reduced alertness. When we speed up and give less focus, this situation can happen. For instance, when we focus on the severity of accidents in rainy conditions, the severity level of the accident is high compared with clear conditions. This dataset is not concerned about serious accidents; it focuses on minor accidents. Let's focus on how road types affect accidents and their severity.



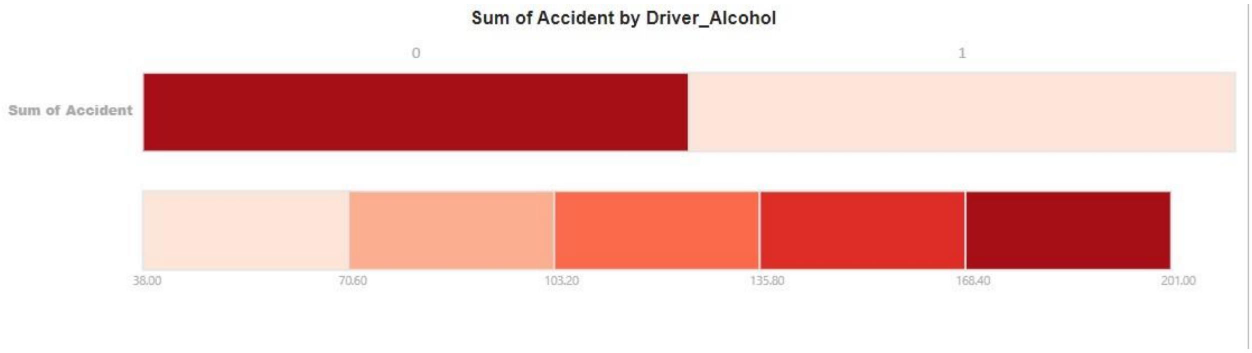
Only weather conditions are impacted by accidents, then. What about the road types? The above graph indicates how the types of roads impact the accidents and their severity. City roads and highways dominate both in accident frequency and severity. High traffic density and unpredictable driver behaviour create a chaotic driving environment, leading to accidents on city roads. High speeds and fatigue during long journeys contribute to more accidents on highways. When focusing on the severity, more damage is caused by highway accidents. The lowest records are indicated from mountain roads because drivers are more focused due to narrow lanes and sharp turns. This shows that city roads and highways require improved traffic management and safety because drivers are losing their intention and focus due to overconfidence. The accident severity is also high in those road types because the number of accidents increases. So far, highway roads have caused more severe accidents under clear weather conditions.



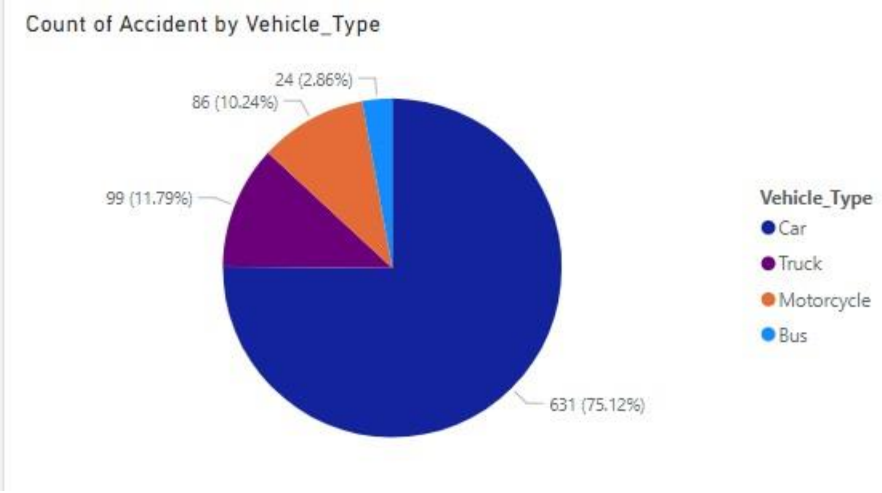
What will happen if inexperienced drivers come onto the road? Will it depend on the road types or drivers' experience? The above graph will show answers to all questions. This graph indicates that accidents are high when road conditions are dry. Typically, we expect more accidents under road conditions like non-constructed, slippery, icy situations. Still, the graph shows us another side of the road conditions because when the road is dry, we do not consider the situation and take risks while driving. This will lead to higher accident records when compared with other road conditions. When we give our attention to the average driver's experience line, it will get us too deep into the visualization because the accident count and the severity of accidents are high at the least average driver experience under dry road conditions. The lowest average driver experience is 35. The accident counts also went high when compared with the situation in which the driver's experience was low. These factors indicate not only the accidents impact outer factors, but also the drivers' behaviour impacts the accidents.



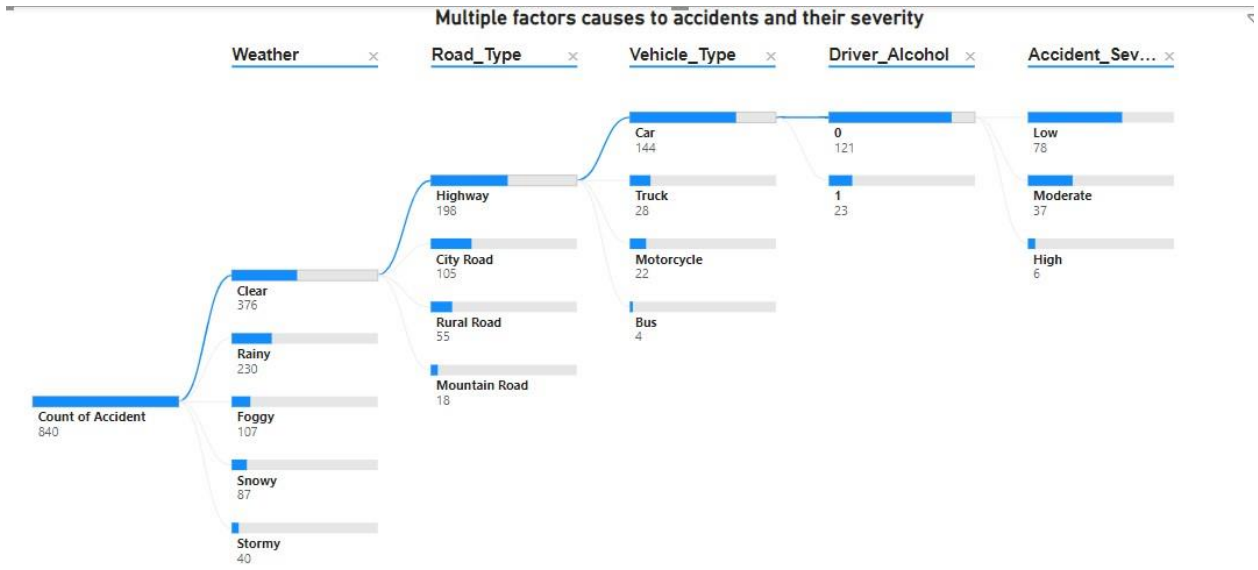
Do you think only the type of roads and weather conditions are impacted by road accidents? Then, look at what we found. Accidents peak during the afternoon, with heavy traffic volume and school pick-ups. This can happen when office workers are going out for their lunch break at this period and must return to work. This can be an issue for accidents in the afternoon period. This dataset shows us the potential factors that can happen in accidents due to the time of the day; everyone knows nighttime has a higher probability of accidents, but this shows the afternoon period has a higher probability of accidents. While decreasing the probability of accidents at night, we should also decrease the accident rate in the afternoon. Only the time of the day decides the accidents and the severity, so what about the drivers' alcohol consumption? Isn't it an issue?



What will happen if the driver gets some drinks over the limit? The accidents have occurred not only because of natural factors but also because factors like alcohol consumption impact the accidents. The heatmap clearly illustrates the sum of accidents based on the consumption of alcohol, which is represented by (0) and non-consumption (1). The drivers who consumed alcohol are associated with a high total number of accidents, as represented by the deep red section from the above heatmap. The sober drivers show a lower total number of accidents, indicated by the lighter section. Alcohol consumption will affect the driver's ability to make quick decisions and respond to road threats while driving on the roads. When alcohol consumption goes over the limit, drivers try to take risks and engage in reckless driving.



The pie chart above shows how the number of accidents depends on the type of vehicle used. Cars dominate accidents by representing 75% probability from all accidents in the dataset. The main reason for these figures is that cars are the most used transportation type worldwide. “Among the countries selected here, the United States has the highest share of personal cars in the commuting population, cited by 75 percent of those surveyed.”(Martin Armstrong, Sep 19 2022) The above statement will confirm the figures shown in the pie chart. When the probability of using goes high, the probability of accidents is also high. In my view, the number of car accidents is high because the usage of cars is higher than that of other types.



Our analysis reveals a straightforward narrative regarding road accidents followed by road types, rod conditions, weather conditions, vehicle types, and drivers' behaviour. The above infographic tells us the summary of the story.

This analysis leads us to the causes of accidents, like weather and road type. Also, if drivers follow good behaviour under any conditions, it will help us avoid accidents and accident severity.

The road ahead isn't just about data points and graphs; it's about taking responsibility and awareness of their self-lives and others and ensuring safety. Even policymakers can take some steps to avoid accident records and accident severity to make a safer road in future.

DRIVE SAFE, KEEP RESPONSIBILITIES IN MIND, AND MAKE EVERY JOURNEY SAFE AND COUNT