DATASET REQUIREMENTS:

Link to dataset: https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

- This dataset on car accidents in the USA includes data from all 49 states. Data on accidents are gathered between February 2016 and December 2021.
- The dataset contains 47 columns with different data types, including integer, float, varchar, and Boolean, and more than 2.8 million records. The organized dataset is 1.15 GB in size.

BUSINESS UNDERSTANDING:

1. **Why has this data been gathered?**
   The purpose of collecting and analyzing this data is to better understand and evaluate traffic accidents in the United States. According to the figures, there were almost 3 million traffic accidents in the US between 2016 and 2020. The information includes a wide range of details, including the location, date, time, weather, accident severity, and the number of cars involved, among others.
   To better prevent and manage accidents, increase road safety, and decrease the number of fatalities and injuries on American roads, researchers, government organizations, and private businesses can use this information to gain insights into the causes and patterns of traffic accidents in the United States.

2. **What can be done with the data? What can we achieve?**
   The US accidents dataset provides valuable information that can be used to improve road safety. It helps us identify high-risk areas, common accident types, and injury severity. This information can be used to implement safety measures, improve emergency response times, and allocate appropriate resources for different types of accidents. Additionally, the dataset can be used to create predictive models for accidents occurring at different locations, times of day, and weather conditions. Policymakers and researchers can also use this data to develop evidence-based policies to enhance road safety.

3. **What are some of the goals/targets we have regarding the business that we can achieve by investigating this data?**
   The US Accidents dataset can help us identify accident-prone areas and improve road safety by implementing measures such as improving road design, installing traffic signals, and increasing police patrols.
   Insurance companies can use the data to identify high-risk areas and adjust their premiums accordingly to reduce the number of claims they receive and keep their costs down.

4. **What insightful information can this data provide us that can be used to improve the business?**
   The US Accidents dataset can help businesses identify areas where accidents are more likely to occur, which can inform their business strategy and marketing efforts.
   Predictive models based on the data can help businesses make informed decisions on resource allocation and risk mitigation strategies, resulting in better safety measures.

5. **Why are we studying this data?**
   The US Accidents dataset can provide valuable insights into the trends and patterns of traffic accidents in the US, which can help identify areas for improvement in road safety.
   Studying the characteristics of accidents, such as contributing factors, types of vehicles, the severity of injuries, and time and location, can help develop strategies to mitigate the impact of accidents on individuals, communities, and the economy.
   The data can also be used by insurance companies, law enforcement agencies, and transportation authorities to identify high-risk areas and populations and develop targeted interventions to prevent accidents or respond to them more effectively.

6. **Are there any problems in our business (based on the given data)?**
   The data suggests that accidents are more likely to occur on certain days of the week or times of the day. This could be a problem for the business if they operate during these high-risk periods, as it could increase the likelihood of accidents for their employees or customers.
   The data shows that certain road types, such as highways or intersections, have a higher frequency of accidents. This could be a problem for the business if they operate in areas with high accident rates, as it could increase the risk of accidents for their employees or customers.

7. **Can we find any solutions to these problems by studying this data?**
   Yes, studying this data can help identify potential solutions to the problems highlighted. For example, businesses can identify accident-prone areas and take steps to improve road safety by improving road design, installing traffic signals, increasing police patrols, or warning drivers of high-risk areas. They can adjust their operations to avoid peak accident times or monitor weather forecasts and adjust their operations accordingly. The US Accidents dataset provides valuable information to develop evidence-based policies and interventions aimed at improving road safety, reducing costs, and increasing customer satisfaction.

8. **What are some of the things we can optimize/improve in our business by studying this data?**
   Studying the US Accidents dataset can help businesses optimize their operations and reduce costs by identifying accident causes and factors, and prioritizing resource allocation. This can lead to lower insurance costs and greater profitability.

9. **What are the main causes of accidents in the United States, and how do they vary by state, time of day, weather conditions, and road type?**
   The US accidents dataset provides information on the type of collision, contributing factors, and road features. By analyzing this information, it is possible to identify the most common causes of accidents in the US and how they differ across different dimensions like states, time of day, weather conditions, and road types.

10. **What is the overall trend of accidents in the United States over time, and are there any patterns or anomalies in the data that can be identified?**
    The US accidents dataset contains details on the occurrence and seriousness of accidents. By analyzing this data, one can find trends and regional differences in accident rates over time, such as seasonal or weather patterns. It can also assist in locating locations with high or low accident rates and any outliers that might need specialized safety initiatives.

11. **How do different variables, such as road conditions, visibility, and traffic volume, affect the severity of accidents and the likelihood of fatalities or injuries?**
    The US accidents dataset includes statistics on factors including traffic volume, visibility, and road conditions that influence accident severity and the likelihood of fatalities or injuries. Analyzing these factors, combined with accident severity and outcome, can help identify the factors that have the most impact on accidents and propose strategies to improve road safety.

DATA UNDERSTANDING:

- **What information each column of the data contains:**
- **The data types of each column:**

#	Attribute	Description	Data Type
1	ID	This is a unique identifier of the accident record.	VARCHAR
2	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 because of the accident) and 4 indicates a significant impact on traffic (i.e., long delay).	INTEGER
3	Start_Time	Shows start time of the accident in local time zone.	DATETIME
4	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.	DATETIME
5	Start_Lat	Shows latitude in GPS coordinate of the start point.	FLOAT
6	Start_Lng	Shows longitude in GPS coordinate of the start point.	FLOAT
7	End_Lat	Shows latitude in GPS coordinate of the end point.	FLOAT
8	End_Lng	Shows longitude in GPS coordinate of the end point.	FLOAT
9	Distance(mi)	The length of the road extent affected by the accident.	FLOAT
10	Description	Shows natural language description of the accident.	VARCHAR
11	Number	Shows the street number in address field.	VARCHAR
12	Street	Shows the street name in address field.	VARCHAR
13	Side	Shows the relative side of the street (Right/Left) in address field.	VARCHAR
14	City	Shows the city in address field.	VARCHAR
15	County	Shows the county in address field.	VARCHAR
16	State	Shows the state in address field.	VARCHAR
17	Zipcode	Shows the zipcode in address field.	VARCHAR
18	Country	Shows the country in address field.	VARCHAR
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	VARCHAR

20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	VARCHAR
21	Weather_Timestamp	Shows the timestamp of weather observation record (in local time).	DATETIME
22	Temperature(F)	Shows the temperature (in Fahrenheit).	FLOAT
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).	FLOAT
24	Humidity (%)	Shows the humidity (in percentage).	FLOAT
25	Pressure(in)	Shows the air pressure (in inches).	FLOAT
26	Visibility(mi)	Shows visibility (in miles).	FLOAT
27	Wind_Direction	Shows wind direction.	VARCHAR
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).	FLOAT
29	Precipitation(in)	Shows precipitation amount in inches if there is any.	FLOAT
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	VARCHAR
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.	BOOLEAN
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	BOOLEAN
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.	BOOLEAN
34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	BOOLEAN
35	Junction	A POI annotation which indicates presence of junction in a nearby location.	BOOLEAN
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	BOOLEAN
37	Railway	A POI annotation which indicates presence of railway in a nearby location.	BOOLEAN
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	BOOLEAN
39	Station	A POI annotation which indicates presence of station in a nearby location.	BOOLEAN

40	Stop	A POI annotation which indicates presence of stop in a nearby location.	BOOLEAN
41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	BOOLEAN
42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	BOOLEAN
43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	BOOLEAN
44	Sunrise_Sunset	Shows the period of day (i.e., day or night) based on sunrise/sunset.	VARCHAR
45	Civil_Twilight	Shows the period of day (i.e., day or night) based on civil twilight.	VARCHAR
46	Nautical_Twilight	Shows the period of day (i.e., day or night) based on nautical twilight.	VARCHAR
47	Astronomical_Twilight	Shows the period of day (i.e., day or night) based on astronomical twilight.	VARCHAR

- **What are some of the values each column contains:**
  - **Describe the values, scales, the range of the data:**

Values: The dataset contains various values related to the accidents such as the time and date of the accident, the severity of the accident, the weather condition during the accident, the location of the accident, the type of junction where the accident occurred, and many others.
Scales: The scales of the data vary depending on the column. For example, the 'Severity' column has a scale from 1 to 4, where 1 represents the least severe accidents and 4 represents the most severe accidents. The 'Temperature(F)' column has a scale from -89.0 to 1049.0 degrees Fahrenheit, representing the temperature at the time of the accident.
Range: The range of the data also varies depending on the column. For example, the 'Distance(mi)' column has a range from 0.0 to 333.630005 miles, representing the distance of the accident from the starting point. The 'Pressure (in)' column has a range from 0.0 to 31.1 inches of mercury, representing the atmospheric pressure at the time of the accident.

- **Verify the data quality:**
  - **Verify the quality of the name of the columns:**
  - **Do you need to change any of the column names? Propose proper column names if the names do not look good to you:**

Here are my proposed new column names:
Change "Start_Lat" to "Start_Latitude": This will make it clear that this column represents the latitude of the start point of the accident.
Change "Start_Lng" to "Start_Longitude": This will make it clear that this column represents the longitude of the start point of the accident.
Change "End_Lat" to "End_Latitude": This will make it clear that this column represents the latitude of the end point of the accident.
Change "End_Lng" to "End_Longitude": This will make it clear that this column represents the longitude of the end point of the accident.

- **Are there any missing values? If yes, then what columns and what percentage?**
- **Are there any duplicate data?**

#	Attribute	No of missing values	% Of Missing Values
1	ID	0	0
2	Severity	0	0
3	Start_Time	0	0
4	End_Time	0	0
5	Start_Lat	0	0
6	Start_Lng	0	0
7	End_Lat	0	0
8	End_Lng	0	0
9	Distance(mi)	0	0
10	Description	0	0
11	Number	1743911	61.29
12	Street	0	0
13	Side	0	0
14	City	0	0
15	County	0	0
16	State	0	0
17	Zipcode	0	0
18	Country	0	0
19	Timezone	0	0
20	Airport_Code	0	0
21	Weather_Timestamp	0	0

22	Temperature(F)	69274	2.43
23	Wind_Chill(F)	469643	16.50
24	Humidity (%)	73092	2.56
25	Pressure(in)	59200	2.08
26	Visibility(mi)	70546	2.47
27	Wind_Direction	0	0
28	Wind_Speed(mph)	157944	5.55
29	Precipitation(in)	549458	19.31
30	Weather_Condition	0	0
31	Amenity	0	0
32	Bump	0	0
33	Crossing	0	0
34	Give_Way	0	0
35	Junction	0	0
36	No_Exit	0	0
37	Railway	0	0
38	Roundabout	0	0
39	Station	0	0
40	Stop	0	0
41	Traffic_Calming	0	0
42	Traffic_Signal	0	0
43	Turning_Loop	0	0
44	Sunrise_Sunset	0	0

45	Civil_Twilight	0	0
46	Nautical_Twilight	0	0
47	Astronomical_Twilight	0	0

Yes, there are duplicate records in the dataset.

- **Do you believe there are outliers in the data:**

  It is possible that there are outliers in the numerical columns such as Distance (mi), Temperature (F), Humidity (%), Pressure (in), Visibility (mi), and Wind_Speed (mph).

- **Provide simple statistics of the data for each column:**

Column	Maximum	Minimum	Mean	Std Dev	Median
Distance(mi)	155.86	0	0.70	1.56	0.24
Severity	4	1	2.13	0.47	2
Wind_Speed(mph)	1087	0	7.39	5.52	7
pressure(in)	58.9	0	29.47	1.04	29.82

Based on the information provided, here are some inferences that could be made:

Distance(mi): The data shows that the maximum distance traveled is 155.86 miles, while the minimum distance is 0.0 miles. The mean distance is 0.70 miles, which suggests that most of the data points are closer to the minimum value. The standard deviation of 1.56 indicates that the data has a relatively large range, with some data points being much further away than others.

Severity: The severity data ranges from 1 to 4, with a mean severity of 2.13. The median severity is 2, which indicates that most of the data points fall into the lower end of the severity scale. The low standard deviation of 0.47 suggests that the severity values are tightly clustered around the mean.

Wind_Speed(mph): The maximum wind speed recorded is 1087 mph, which is an unusually high value and could be an outlier. The mean wind speed is 7.39 mph, and the median is 7 mph, which indicates that most of the data points fall into a relatively narrow range. The standard deviation of 5.52 indicates that there is some variability in the data, with some wind speeds being much higher or lower than the mean.

Pressure(in): The maximum pressure recorded is 58.9 inches, while the minimum is 0 inches. The mean pressure is 29.47 inches, which is close to the standard atmospheric pressure at sea level. The median pressure is 29.82 inches, which suggests that most of the data points fall into a range close to the mean. The low standard deviation of 1.04 indicates that the pressure values are tightly clustered around the mean.

Overall, the data suggest that the severity of the events recorded is mostly low to moderate, with some outliers in wind speed and pressure. The distance data has a wide range, indicating that some events occurred very far away from the observation point. The wind speed and pressure data show some variability, but most data points are close to the mean values.

- **Try to understand the relationships between the columns of the data:**
  - **What relationships can you find between the columns?**

Relationships that could be explored within the US Accidents dataset:

Time of Day and Severity: It's possible that accidents occur more frequently during certain times of the day or night, and that the severity of those accidents is also affected by the time of day. By analyzing the Hour column and comparing it to the Severity column, it may be possible to identify patterns such as more severe accidents occurring during rush hour traffic.

Weather Condition and Accident Severity: It's likely that the weather conditions at the time of an accident can have an impact on the severity of the accident. By analyzing the Weather_Condition column and comparing it to the Severity column, it may be possible to identify patterns such as more severe accidents occurring during rainy or snowy conditions.

Location and Accident Type: It's possible that certain types of accidents are more common in certain geographic locations. By analyzing the Latitude, Longitude, and Accident_Type columns, it may be possible to identify patterns such as more vehicle collisions occurring in urban areas, or more animal-related accidents occurring in rural areas.

Distance and Accident Severity: It's possible that the distance between the accident location and the nearest traffic signal, stop sign, or other traffic control device can affect the severity of the accident. By analyzing the Distance column and comparing it to the Severity column, it may be possible to identify patterns such as more severe accidents occurring in locations with fewer traffic control devices.

- **Is there a connection between the columns? Describe.**

The columns Start_Lat and Start_Lng provide the geographical coordinates of the accident location. Similarly, the columns Weather_Condition, Temperature(F), Visibility(mi), and Wind_Speed(mph) provide details about the weather conditions at the time of the accident. And for weather conditions and the severity of accidents. If certain types of weather conditions, such as heavy rain or snow, tend to be associated with more severe accidents, then there would be a connection between these two columns.