Applied Analytics Project

Analyzing US Accident Data to Predict High-Risk Areas and Times in Massachusetts

Major: Applied Analytics

Name: Gefan Wang, Chenhe Shi, Tianchen Liu

Date: 02.02.2025

1. What is the target variable and why?

For our group project, the target variable is Severity. This variable represents the severity level of each accident, a number between 1 and 4, where 1 indicates the least impact on traffic. Focusing on this target helps us identify patterns and factors contributing to high-severity incidents, which are critical for improving public safety. Severity level would also be crucial for policymakers and emergency responders since it helps them prioritize and allocate resources more effectively to areas or conditions that frequently result in severe accidents. Severity prediction can also inform infrastructure improvements, traffic regulation, and safety campaigns. For instance, areas prone to severe accidents can have better-equipped ambulances, quicker response teams, or advanced trauma centers nearby, reducing fatalities and long-term injuries.

2. What are the predictors and why?

Predictors are variables that can influence the target variable (Severity). Based on the dataset, we have separated the predictors into five categories.

The first category is Weather Conditions, which include variables such as temperature, humidity, visibility, precipitation, and wind speed. Weather significantly impacts road safety by affecting vehicle control, braking distance, and driver visibility. For example, heavy rain or fog can reduce visibility, increasing the likelihood of severe accidents, while icy roads can cause vehicles to skid, resulting in more serious collisions.

The second category is Geographic Information, which consists of latitude, longitude, city, county, and state. These variables can help identify areas with higher accident risks due to factors like poor road infrastructure, high population density, or challenging terrain. Geographic insights can aid in understanding regional trends and targeting specific locations for interventions.

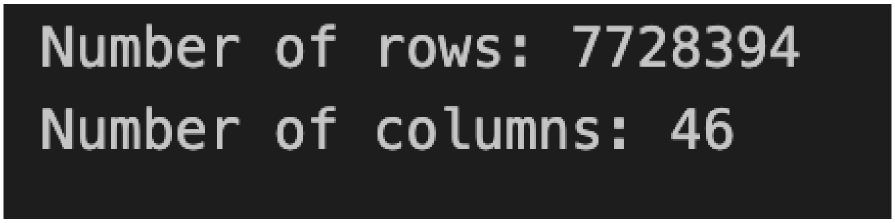
The third category is Infrastructure Features, which include attributes such as the presence of bumps, crossings, junctions, railway tracks, stops, and traffic signals. These features directly influence the frequency and severity of accidents by controlling traffic flow and providing safeguards. For example, intersections with inadequate signage or poorly maintained railway crossings can increase the likelihood of severe accidents.

The fourth category is Time Features, including start time, end time, timezone, and twilight conditions. The timing of accidents plays a critical role in determining severity. Accidents at night or during twilight, when visibility is reduced, may be more severe. Similarly, accidents during peak traffic hours often involve multiple vehicles, potentially increasing their severity.

The fifth category is Description, which provides textual data containing additional context about accident causes, road conditions, and involved parties. This field can be analyzed using Natural Language Processing (NLP) techniques to extract meaningful patterns and insights, such as mentions of distracted driving, speeding, or hazardous conditions.

Together, these categories of predictors provide a comprehensive framework for understanding the factors that influence accident severity, enabling targeted interventions and predictive modeling for improved road safety.

3. Exploration of the dataset: definition of variables, data types, general dataset stats: count of rows, count of columns, etc.



```
Number of rows: 7728394
Number of columns: 46
```

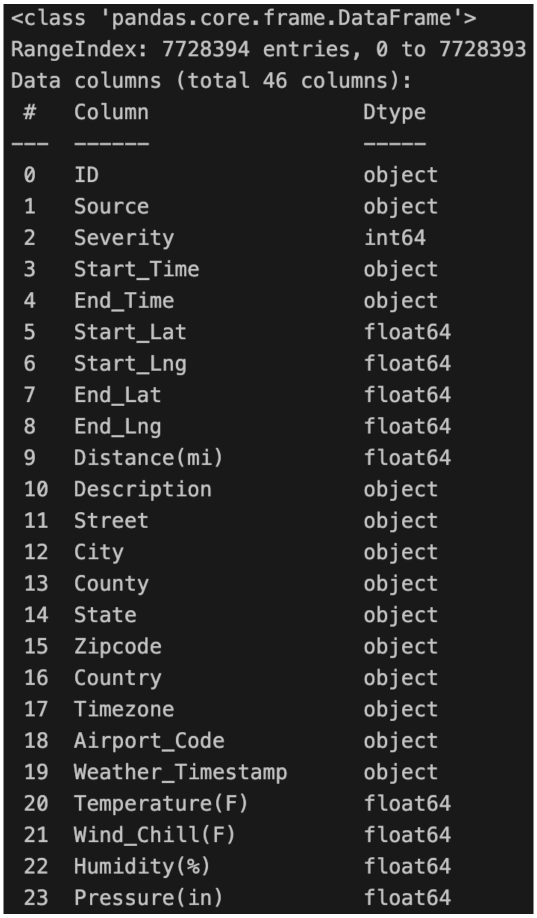
The dataset contains 7,728,394 rows and 46 columns, indicating a large-scale dataset with extensive accident records. Since we only focused on accidents that happened in Massachusetts, the rows are reduced to 62,000. Table of variables with their definition are shown below:

- ID - This is a unique identifier of the accident record.
- 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 as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long 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.
- Start_Lat - Shows latitude in GPS coordinate of the start point.
- Start_Lng - Shows longitude in GPS coordinate of the start point.

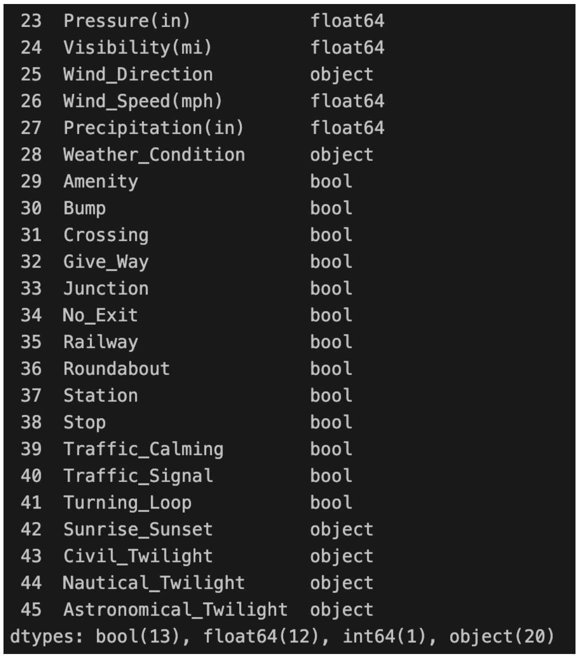
- End_Lat - Shows latitude in GPS coordinate of the end point.
- End_Lng - Shows longitude in GPS coordinate of the end point.
- Distance(mi) - The length of the road extent affected by the accident.
- Description - Shows natural language description of the accident.
- Number - Shows the street number 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.
- Zipcode - Shows the zipcode in address record.
- Country - Shows the country in address record.
- Timezone - Shows timezone based on the location of the accident (eastern, central, etc.).
- Airport_Code - Denotes an airport-based weather station which is the closest one to location of the accident.
- Weather_Timestamp - Shows the time-stamp of weather observation record (in local time).
- Temperature(F) - Shows the temperature (in Fahrenheit).
- Wind_Chill(F) - Shows the wind chill (in Fahrenheit).
- Humidity(%) - Shows the humidity (in percentage).
- Pressure(in) - Shows the air pressure (in inches).
- Visibility(mi) - Shows visibility (in miles).
- Wind_Direction - Shows wind direction.
- Wind_Speed(mph) - Shows wind speed (in miles per hour).
- Precipitation(in) - Shows precipitation amount in inches, if there is any.
- Weather_Condition - Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
- Amenity - A POI annotation which indicates presence of amenity in a nearby location.
- Bump - A POI annotation which indicates presence of speed bump or hump in a nearby location.
- Crossing - A POI annotation which indicates presence of crossing in a nearby location.
- Give_Way - A POI annotation which indicates presence of give_way in a nearby location.
- Junction - A POI annotation which indicates presence of junction in a nearby location.

- No_Exit - A POI annotation which indicates presence of junction in a nearby location.
- Railway - A POI annotation which indicates presence of railway in a nearby location.
- Roundabout - A POI annotation which indicates presence of roundabout in a nearby location.
- Station - A POI annotation which indicates presence of station in a nearby location.
- Stop - A POI annotation which indicates presence of stop in a nearby location.
- Traffic_Calming - A POI annotation which indicates presence of traffic_calming in a nearby location.
- Traffic_Signal - A POI annotation which indicates presence of traffic_signal in a nearby location.
- Turning_Loop - A POI annotation which indicates presence of turning_loop in a nearby location.
- Sunrise_Sunset - Shows the period of day (i.e. day or night) based on sunrise/sunset.
- Civil_Twilight - Shows the period of day (i.e. day or night) based on civil twilight.
- Nautical_Twilight - Shows the period of day (i.e. day or night) based on nautical twilight.
- Astronomical_Twilight - Shows the period of day (i.e. day or night) based on astronomical twilight.

B. Data Types



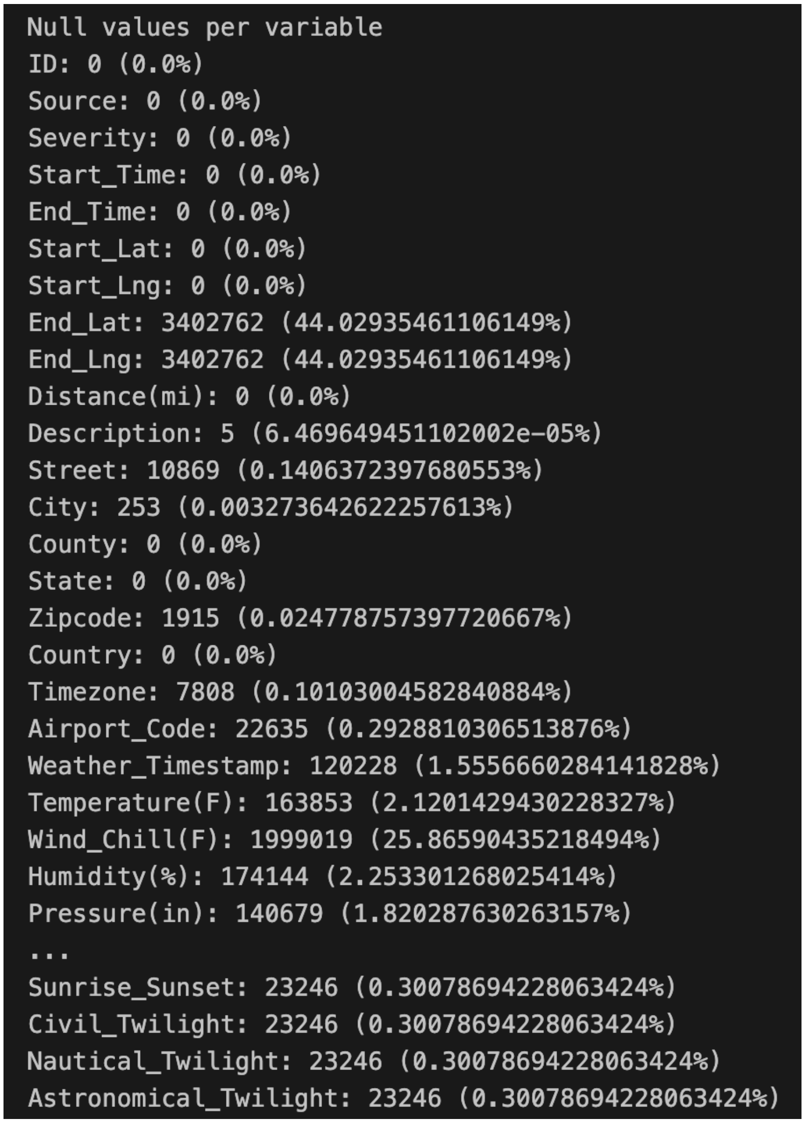
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7728394 entries, 0 to 7728393
Data columns (total 46 columns):
 #   Column                 Dtype
---  ------                 -----
 0   ID                     object
 1   Source                 object
 2   Severity               int64
 3   Start_Time             object
 4   End_Time               object
 5   Start_Lat              float64
 6   Start_Lng              float64
 7   End_Lat                float64
 8   End_Lng                float64
 9   Distance(mi)           float64
 10  Description            object
 11  Street                 object
 12  City                   object
 13  County                 object
 14  State                  object
 15  Zipcode                object
 16  Country                object
 17  Timezone               object
 18  Airport_Code           object
 19  Weather_Timestamp      object
 20  Temperature(F)         float64
 21  Wind_Chill(F)          float64
 22  Humidity(%)            float64
 23  Pressure(in)           float64
```



```
 23  Pressure(in)           float64
 24  Visibility(mi)         float64
 25  Wind_Direction         object
 26  Wind_Speed(mph)        float64
 27  Precipitation(in)      float64
 28  Weather_Condition      object
 29  Amenity                bool
 30  Bump                   bool
 31  Crossing               bool
 32  Give_Way               bool
 33  Junction               bool
 34  No_Exit                bool
 35  Railway                bool
 36  Roundabout             bool
 37  Station                bool
 38  Stop                   bool
 39  Traffic_Calming        bool
 40  Traffic_Signal         bool
 41  Turning_Loop           bool
 42  Sunrise_Sunset         object
 43  Civil_Twilight         object
 44  Nautical_Twilight      object
 45  Astronomical_Twilight  object
dtypes: bool(13), float64(12), int64(1), object(20)
```

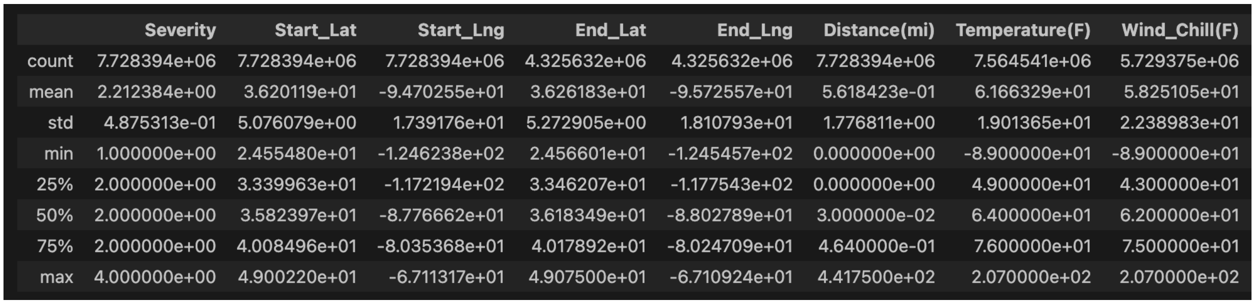
The dataset consists of 46 columns with a mix of 13 boolean, 12 float, 1 integer, and 20 object data types. Numerical columns like temperature and visibility provide quantitative insights, while object columns like weather conditions and descriptions may require encoding or NLP processing.

C. General Dataset Stats



```
Null values per variable
ID: 0 (0.0%)
Source: 0 (0.0%)
Severity: 0 (0.0%)
Start_Time: 0 (0.0%)
End_Time: 0 (0.0%)
Start_Lat: 0 (0.0%)
Start_Lng: 0 (0.0%)
End_Lat: 3402762 (44.02935461106149%)
End_Lng: 3402762 (44.02935461106149%)
Distance(mi): 0 (0.0%)
Description: 5 (6.469649451102002e-05%)
Street: 10869 (0.1406372397680553%)
City: 253 (0.003273642622257613%)
County: 0 (0.0%)
State: 0 (0.0%)
Zipcode: 1915 (0.024778757397720667%)
Country: 0 (0.0%)
Timezone: 7808 (0.10103004582840884%)
Airport_Code: 22635 (0.2928810306513876%)
Weather_Timestamp: 120228 (1.5556660284141828%)
Temperature(F): 163853 (2.1201429430228327%)
Wind_Chill(F): 1999019 (25.86590435218494%)
Humidity(%): 174144 (2.253301268025414%)
Pressure(in): 140679 (1.820287630263157%)
...
Sunrise_Sunset: 23246 (0.30078694228063424%)
Civil_Twilight: 23246 (0.30078694228063424%)
Nautical_Twilight: 23246 (0.30078694228063424%)
Astronomical_Twilight: 23246 (0.30078694228063424%)
```

The missing values analysis shows that some columns, such as **End_Lat** and **End_Lng**, have a high percentage of missing data (44%), while others like **Wind_Chill(F)** (25%) and **Temperature(F)** (2%) have moderate missing values. Certain categorical variables like **Street** and **Timezone** also have small percentages of missing data. Handling missing values through imputation or removal will be essential for ensuring data quality and model accuracy.



	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)
count	7.728394e+06	7.728394e+06	7.728394e+06	4.325632e+06	4.325632e+06	7.728394e+06	7.564541e+06	5.729375e+06
mean	2.212384e+00	3.620119e+01	-9.470255e+01	3.626183e+01	-9.572557e+01	5.618423e-01	6.166329e+01	5.825105e+01
std	4.875313e-01	5.076079e+00	1.739176e+01	5.272905e+00	1.810793e+01	1.776811e+00	1.901365e+01	2.238983e+01
min	1.000000e+00	2.455480e+01	-1.246238e+02	2.456601e+01	-1.245457e+02	0.000000e+00	-8.900000e+01	-8.900000e+01
25%	2.000000e+00	3.339963e+01	-1.172194e+02	3.346207e+01	-1.177543e+02	0.000000e+00	4.900000e+01	4.300000e+01
50%	2.000000e+00	3.582397e+01	-8.776662e+01	3.618349e+01	-8.802789e+01	3.000000e-02	6.400000e+01	6.200000e+01
75%	2.000000e+00	4.008496e+01	-8.035368e+01	4.017892e+01	-8.024709e+01	4.640000e-01	7.600000e+01	7.500000e+01
max	4.000000e+00	4.900220e+01	-6.711317e+01	4.907500e+01	-6.710924e+01	4.417500e+02	2.070000e+02	2.070000e+02

Above are statistical graphs of the dataset. The summary statistics indicate that most accidents have a severity level around 2, with localized incidents (median distance of 0.03 miles). Additionally, extreme values in temperature and wind chill (-89°F to 207°F) suggest potential outliers that may need cleaning for accurate analysis.

Week2 Code_EDA

February 2, 2025

1 Week 2. Basic EDA

[2]:

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

[3]:

```
#import data
accident_data = pd.read_csv('/Users/wanggefan/Desktop/2025 Spring/ Applied␣
 ↪Analytics Project/US_Accidents_March23.csv')
```

[4]:

```
#look at datatype
accident_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7728394 entries, 0 to 7728393
Data columns (total 46 columns):
 #   Column                 Dtype
---  ------                 -----
 0   ID                     object
 1   Source                 object
 2   Severity               int64
 3   Start_Time             object
 4   End_Time               object
 5   Start_Lat              float64
 6   Start_Lng              float64
 7   End_Lat                float64
 8   End_Lng                float64
 9   Distance(mi)           float64
 10  Description            object
 11  Street                 object
 12  City                   object
 13  County                 object
 14  State                  object
 15  Zipcode                object
 16  Country                object
 17  Timezone               object
 18  Airport_Code           object
 19  Weather_Timestamp      object
 20  Temperature(F)         float64
 21  Wind_Chill(F)          float64
 22  Humidity(%)            float64
 23  Pressure(in)           float64
 24  Visibility(mi)         float64
 25  Wind_Direction         object
 26  Wind_Speed(mph)        float64
 27  Precipitation(in)      float64
 28  Weather_Condition      object
 29  Amenity                bool
 30  Bump                   bool
 31  Crossing               bool
 32  Give_Way               bool
 33  Junction               bool
 34  No_Exit                bool
 35  Railway                bool
 36  Roundabout             bool
 37  Station                bool
 38  Stop                   bool
 39  Traffic_Calming        bool
 40  Traffic_Signal         bool
 41  Turning_Loop           bool
 42  Sunrise_Sunset         object
 43  Civil_Twilight         object
 44  Nautical_Twilight      object
 45  Astronomical_Twilight  object
dtypes: bool(13), float64(12), int64(1), object(20)
memory usage: 2.0+ GB
```

[5]:

```
#print number and percentage of null entries per variable
print('Null values per variable')
for column in accident_data.columns:
    print('{}: {} ({}%)'.format(column,pd.isnull(accident_data[column]).
 ↪sum(),(pd.isnull(accident_data[column]).sum()/len(accident_data))*100))
```

```
Null values per variable
ID: 0 (0.0%)
Source: 0 (0.0%)
Severity: 0 (0.0%)
Start_Time: 0 (0.0%)
End_Time: 0 (0.0%)
Start_Lat: 0 (0.0%)
Start_Lng: 0 (0.0%)
End_Lat: 3402762 (44.02935461106149%)
End_Lng: 3402762 (44.02935461106149%)
```

```
Distance(mi): 0 (0.0%)
Description: 5 (6.469649451102002e-05%)
Street: 10869 (0.1406372397680553%)
City: 253 (0.003273642622257613%)
County: 0 (0.0%)
State: 0 (0.0%)
Zipcode: 1915 (0.024778757397720667%)
Country: 0 (0.0%)
Timezone: 7808 (0.10103004582840884%)
Airport_Code: 22635 (0.2928810306513876%)
Weather_Timestamp: 120228 (1.5556660284141828%)
Temperature(F): 163853 (2.1201429430228327%)
Wind_Chill(F): 1999019 (25.86590435218494%)
Humidity(%): 174144 (2.253301268025414%)
Pressure(in): 140679 (1.820287630263157%)
Visibility(mi): 177098 (2.291523956982524%)
Wind_Direction: 175206 (2.2670428034595544%)
Wind_Speed(mph): 571233 (7.391354529802699%)
Precipitation(in): 2203586 (28.512857910712107%)
Weather_Condition: 173459 (2.244437848277404%)
Amenity: 0 (0.0%)
Bump: 0 (0.0%)
Crossing: 0 (0.0%)
Give_Way: 0 (0.0%)
Junction: 0 (0.0%)
No_Exit: 0 (0.0%)
Railway: 0 (0.0%)
Roundabout: 0 (0.0%)
Station: 0 (0.0%)
Stop: 0 (0.0%)
Traffic_Calming: 0 (0.0%)
Traffic_Signal: 0 (0.0%)
Turning_Loop: 0 (0.0%)
Sunrise_Sunset: 23246 (0.30078694228063424%)
Civil_Twilight: 23246 (0.30078694228063424%)
Nautical_Twilight: 23246 (0.30078694228063424%)
Astronomical_Twilight: 23246 (0.30078694228063424%)
```

[6]:
```
#look at distribution of data
accident_data.describe()
```

[6]:
```
           Severity     Start_Lat     Start_Lng       End_Lat       End_Lng  \
count  7.728394e+06  7.728394e+06  7.728394e+06  4.325632e+06  4.325632e+06
mean   2.212384e+00  3.620119e+01 -9.470255e+01  3.626183e+01 -9.572557e+01
std    4.875313e-01  5.076079e+00  1.739176e+01  5.272905e+00  1.810793e+01
min    1.000000e+00  2.455480e+01 -1.246238e+02  2.456601e+01 -1.245457e+02
25%    2.000000e+00  3.339963e+01 -1.172194e+02  3.346207e+01 -1.177543e+02
```

```
50%    2.000000e+00  3.582397e+01 -8.776662e+01  3.618349e+01 -8.802789e+01
75%    2.000000e+00  4.008496e+01 -8.035368e+01  4.017892e+01 -8.024709e+01
max    4.000000e+00  4.900220e+01 -6.711317e+01  4.907500e+01 -6.710924e+01

       Distance(mi)  Temperature(F)  Wind_Chill(F)   Humidity(%)  \
count  7.728394e+06    7.564541e+06   5.729375e+06  7.554250e+06
mean   5.618423e-01    6.166329e+01   5.825105e+01  6.483104e+01
std    1.776811e+00    1.901365e+01   2.238983e+01  2.282097e+01
min    0.000000e+00   -8.900000e+01  -8.900000e+01  1.000000e+00
25%    0.000000e+00    4.900000e+01   4.300000e+01  4.800000e+01
50%    3.000000e-02    6.400000e+01   6.200000e+01  6.700000e+01
75%    4.640000e-01    7.600000e+01   7.500000e+01  8.400000e+01
max    4.417500e+02    2.070000e+02   2.070000e+02  1.000000e+02

       Pressure(in)  Visibility(mi)  Wind_Speed(mph)  Precipitation(in)
count  7.587715e+06    7.551296e+06     7.157161e+06       5.524808e+06
mean   2.953899e+01    9.090376e+00     7.685490e+00       8.407210e-03
std    1.006190e+00    2.688316e+00     5.424983e+00       1.102246e-01
min    0.000000e+00    0.000000e+00     0.000000e+00       0.000000e+00
25%    2.937000e+01    1.000000e+01     4.600000e+00       0.000000e+00
50%    2.986000e+01    1.000000e+01     7.000000e+00       0.000000e+00
75%    3.003000e+01    1.000000e+01     1.040000e+01       0.000000e+00
max    5.863000e+01    1.400000e+02     1.087000e+03       3.647000e+01
```

[9]:

```python
# Get the number of rows and columns
num_rows, num_columns = accident_data.shape

print(f"Number of rows: {num_rows}")
print(f"Number of columns: {num_columns}")
```

```
Number of rows: 7728394
Number of columns: 46
```

[7]:

```python
#look at formatting of entries
accident_data.head()
```

[7]:

```
    ID   Source  Severity           Start_Time             End_Time  \
0  A-1  Source2         3  2016-02-08 05:46:00  2016-02-08 11:00:00
1  A-2  Source2         2  2016-02-08 06:07:59  2016-02-08 06:37:59
2  A-3  Source2         2  2016-02-08 06:49:27  2016-02-08 07:19:27
3  A-4  Source2         3  2016-02-08 07:23:34  2016-02-08 07:53:34
4  A-5  Source2         2  2016-02-08 07:39:07  2016-02-08 08:09:07

   Start_Lat  Start_Lng  End_Lat  End_Lng  Distance(mi)  … Roundabout  \
0  39.865147 -84.058723      NaN      NaN          0.01  …      False
1  39.928059 -82.831184      NaN      NaN          0.01  …      False
2  39.063148 -84.032608      NaN      NaN          0.01  …      False
3  39.747753 -84.205582      NaN      NaN          0.01  …      False
```

```
4  39.627781 -84.188354      NaN      NaN          0.01  …      False

  Station   Stop Traffic_Calming Traffic_Signal Turning_Loop Sunrise_Sunset  \
0   False  False           False          False        False          Night
1   False  False           False          False        False          Night
2   False  False           False           True        False          Night
3   False  False           False          False        False          Night
4   False  False           False           True        False            Day

  Civil_Twilight Nautical_Twilight Astronomical_Twilight
0          Night             Night                 Night
1          Night             Night                   Day
2          Night               Day                   Day
3            Day               Day                   Day
4            Day               Day                   Day

[5 rows x 46 columns]
```

[8]:

```
#looking to see ID format towards end
accident_data.tail()
```

[8]:

```
                ID   Source  Severity           Start_Time  \
7728389  A-7777757  Source1         2  2019-08-23 18:03:25
7728390  A-7777758  Source1         2  2019-08-23 19:11:30
7728391  A-7777759  Source1         2  2019-08-23 19:00:21
7728392  A-7777760  Source1         2  2019-08-23 19:00:21
7728393  A-7777761  Source1         2  2019-08-23 18:52:06

                    End_Time  Start_Lat  Start_Lng   End_Lat    End_Lng  \
7728389  2019-08-23 18:32:01   34.00248 -117.37936  33.99888 -117.37094
7728390  2019-08-23 19:38:23   32.76696 -117.14806  32.76555 -117.15363
7728391  2019-08-23 19:28:49   33.77545 -117.84779  33.77740 -117.85727
7728392  2019-08-23 19:29:42   33.99246 -118.40302  33.98311 -118.39565
7728393  2019-08-23 19:21:31   34.13393 -117.23092  34.13736 -117.23934

         Distance(mi)  … Roundabout Station   Stop Traffic_Calming  \
7728389         0.543  …      False   False  False           False
7728390         0.338  …      False   False  False           False
7728391         0.561  …      False   False  False           False
7728392         0.772  …      False   False  False           False
7728393         0.537  …      False   False  False           False

        Traffic_Signal Turning_Loop Sunrise_Sunset Civil_Twilight  \
7728389          False        False            Day            Day
7728390          False        False            Day            Day
7728391          False        False            Day            Day
7728392          False        False            Day            Day
```

```
7728393          False        False            Day             Day

        Nautical_Twilight Astronomical_Twilight
7728389               Day                   Day
7728390               Day                   Day
7728391               Day                   Day
7728392               Day                   Day
7728393               Day                   Day

[5 rows x 46 columns]
```