Aim

1. To select features relevant for modelling

In [11]:
```python
import os

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

In [2]:
```python
DATA_ROOT = f"../data"
```

In [3]:
```python
df = pd.read_csv(f"{DATA_ROOT}/eda/clean-data.csv")
df.columns
```

Out[3]:
```
Index(['ID', 'Source', 'TMC', 'Severity', 'Start_Time', 'End_Time',
       'Start_Lat', 'Start_Lng', 'Distance(mi)', 'Description', 'Number',
       'Street', 'Side', 'City', 'County', 'State', 'Zipcode', 'Country',
       'Timezone', 'Airport_Code', 'Weather_Timestamp', 'Temperature(F)',
       'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Direction',
       'Wind_Speed(mph)', 'Weather_Condition', 'Amenity', 'Bump', 'Crossin
g',
       'Give_Way', 'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Statio
n',
       'Stop', 'Traffic_Calming', 'Traffic_Signal', 'Turning_Loop',
       'Sunrise_Sunset', 'Civil_Twilight', 'Nautical_Twilight',
       'Astronomical_Twilight'],
      dtype='object')
```

NOTE

- For first iteration we do not need all the features.
- Trying to keep the feature set simple.
- Variables on which EDA was done will be used in modelling.

Initial thoughts on Feature Selection

1. Categorical variables might help in improving information gain, and better model Severity
2. Hence keeping categorical variables

---

1. Traffic Attributes

    - ID (will be removed before modelling),
    - Source,
    - TMC,
    - start_time, end_time, (NOT using these to reduce complexity of problem)
    - start_point, end_point, (NOT using these to reduce complexity of problem)

- distance (varying lengths of tails were observed for different severity in EDA might be a good feature),
- and description (interesting feature engineering could be done using text keeping this)

2. Address Attributes

    - Number (street number in itself might be a noisy variable + it has 64% Nan values which were filled with -1 skipping this variable)
    - Street (This is potential variable time take to featurise it is a bit high keeping it for future improvements),
    - Side (left/right - this variable has distinct distributions for varying severity observed in EDA),
    - City, County, State, (Keeping these categorical feature as it is),
    - Country (as there is only one country this will be a noise variable to the model)
    - Zipcode (these are hierarical in nature and can be good information to the model),

3. Weather Attributes

    - keeping all the weather attributes as these external factors might help us explain severity of accidents
    - good amount of variance was seen across various weather attributes in EDA
    - skipping these variables due to very high nan %
        - Precipitation, Wind_Chill

4. POI variables

    - keeping all POI variables as varying distributions of these variables were observed in bivariate analysis

5. Period of Day variables

    - keeping these variables
    - Sunrise/Sunset, Civil Twilight, Nautical Twilight, and Astronomical Twilight

6. Others

    - Start_Lat, Start_Lng
    - as we are not having End Lat Lng due to 70% nans in them
    - as features like distance between 2 geo-coordinates cannot be calculated
    - zipcode should make up for the information loss due to geo-coordinates
    - this is a tradeoff that is being taken for this assignment

---

The above selection is totally based on observations in EDA.

1. Better feature selection methods on the basis of feature importance could be employed.
2. Doing a Backward or Forward feature selection will be a time consuming process.
3. Hence keeping it for future scope

In [39]:
```
selected_features = [
    "ID",  # will remove this in the end before modelling
    "Source",
```

```python
    "TMC",
    "Start_Time",  # keeping this to sort the dataframe before train test sp
    # "End_Time",
    # "Start_Lat",
    # "Start_Lng",
    "Distance(mi)",
    "Description",
    # "Number",
    # "Street",
    "Side",
    "City",
    "County",
    "State",
    "Zipcode",
    # "Country",
    "Timezone",
    "Airport_Code",
    # "Weather_Timestamp",
    "Temperature(F)",
    "Humidity(%)",
    "Pressure(in)",
    "Visibility(mi)",
    "Wind_Direction",
    "Wind_Speed(mph)",
    "Weather_Condition",
    "Amenity",
    "Bump",
    "Crossing",
    "Give_Way",
    "Junction",
    "No_Exit",
    "Railway",
    "Roundabout",
    "Station",
    "Stop",
    "Traffic_Calming",
    "Traffic_Signal",
    "Turning_Loop",
    "Sunrise_Sunset",
    "Civil_Twilight",
    "Nautical_Twilight",
    "Astronomical_Twilight",
    #
    "Severity",
]
```

In [37]:

```python
os.makedirs(f"{DATA_ROOT}/fselect/")
```

In [40]:

```python
df[selected_features].to_pickle(f"{DATA_ROOT}/fselect/accidents_raw.pkl")
```