# Exploratory Data Analysis

- StudentID: 21900707

- Name: YoungWoo Cho

- 1st Major: AI Convergence & Entrepreneurship

- 2nd Major: Data Science

*"Analyzing the Characteristics of Severity 2 Accidents in Urban Areas to Identify Prevention and Quick Response Measures"*

## 1. Data overview

- Size:
    - 7,728,394 rows and 46 columns
- Number of NA:
    - 30 out of 46 columns do not have NA values
    - ['ID', 'Source', 'Severity', 'Start_Time', 'End_Time', 'Start_Lat', 'Start_Lng', 'Distance(mi)', 'County', 'State', 'Country', 'Amenity', 'Bump', 'Crossing', 'Give_Way', 'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Station', 'Stop', 'Traffic_Calming', 'Traffic_Signal', 'Turning_Loop']
- Choosing Data:
    - Too much data is lost to analyze after NA data is removed. Thus, 16 columns which contain NA value are abandoned.
    - Removing "Source" value:
        - There are three Sources, but unsure where the sources came from.
    - Removing "Start_Lat", "Start_Lng" values:
        - There are NA values in "End_Lat", "End_Lng", so the information value of the starting and ending points decreases. Additionally, start values overlaps in meaning with the "State" column.
    - Removing "County", "Country" values:
        - "County" cannot act as a factor because there are too many unique variables. "Country" is meaningless because there is only one country, the United States.
- Information of Columns:
    - | Name | Type | Range | Description |
      |---|---|---|---|

| Name | Type | Range | Description |
|---|---|---|---|
| ID | Factor | A-1 ~ A-7777757 | unique identifier of the accident record |
| Severity | Factor | 1,2,3,4 | severity of the accident |
| Start_Time | DateTime | 2016-02-08 05:46:00,..., 2019-08-23 18:52:06 | start time of the accident |
| End_Time | DateTime | 2016-02-08 11:00:00,..., 2019-08-23 19:21:31 | end time of the accident |
| Distance(mi) | numeric | 0 ~ 441.75 | length of the road extent affected by the accident in miles |
| State | character | OH,...,CH | state in address field |
| Amenity | bool | T/F | indicates presence of amenity in a nearby location |
| Bump | bool | T/F | indicates presence of Bump in a nearby location |
| Crossing | bool | T/F | indicates presence of Crossing in a nearby location |
| Give_Way | bool | T/F | indicates presence of Give_Way in a nearby location |
| Junction | bool | T/F | indicates presence of Junction in a nearby location |
| No_Exit | bool | T/F | indicates presence of No_Exit in a nearby location |
| Railway | bool | T/F | indicates presence of Railway in a nearby location |
| Roundabout | bool | T/F | indicates presence of Roundabout in a nearby location |
| Station | bool | T/F | indicates presence of Station in a nearby location |
| Stop | bool | T/F | indicates presence of Stop in a nearby location |
| Traffic_Calming | bool | T/F | indicates presence of Traffic_Calming in a nearby location |

| Name | Type | Range | Description |
|---|---|---|---|
| Traffic_Signal | bool | T/F | indicates presence of amenTraffic_Signality in a nearby location |
| Turning_Loop | bool | T/F | indicates presence of Turning_Loop in a nearby location |

Table 1. Information of Selected Columns

# 2. Univariate analysis

## 2.1 Variable 1 - Severity

- Distribution:
  - | Level | Count | Percnetage |
    |---|---|---|
    | 1 | 67,366 | 0.87% |
    | 2 | 6,156,981 | 79.67% |
    | 3 | 1,299,337 | 16.81% |
    | 4 | 204,710 | 2.65% |

    Table 2. Number and Ratio of Accidents by Levels

- Description:
  - Level 2 accounts for the largest proportion at about 80%, followed by level 3, level 4, and level 1.

## 2.2 Variable 2 - Start_Time & End_Time

- Derived Variable:
  - Start_Month: Accident occurrence month information extracted from Start_Time.

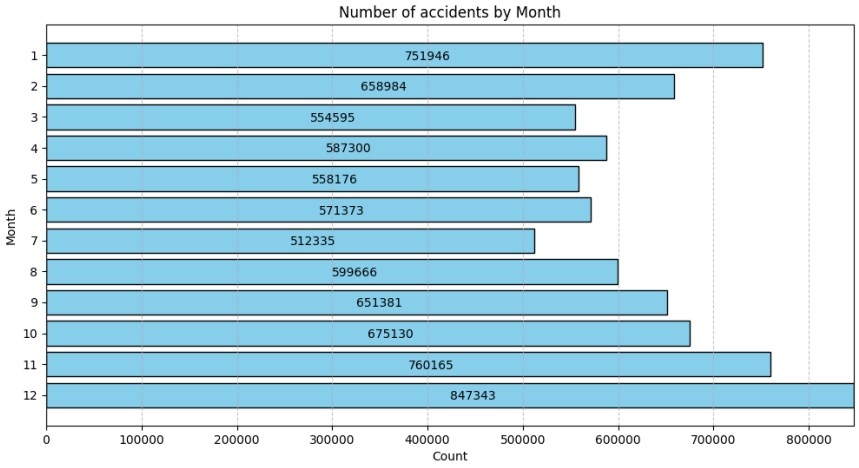


Figure 1. Bar Graph of Monthly Accident Frequency

- Description:
  - Not only can observe the monthly accident trends, but also identify seasonal patterns in accident occurrences. In particular, it is evident that the highest number of accidents occurs in winter, while summer sees the fewest accidents. This suggests a seasonal factor, such as snowfall, contributing to these variations.

- Recovery_Time: Time taken until the accident was resolved. (End_Time) - (Start_Time)
  - Unit: minute
  - Average: 444.42
  - Quantile:
    - 0% - 1.22
    - 25% - 31.50
    - 50% - 74.83
    - 75% - 125.15
    - 100% - 2812939.00

## 2.2 Variable 3 - Distance

- Average: 0.562
- Quantile:
  - 0% - 0.00
  - 25% - 0.00
  - 50% - 0.03
  - 75% - 0.46
  - 100% - 441.75

## 2.2 Variable 4 - State

- Visualization:

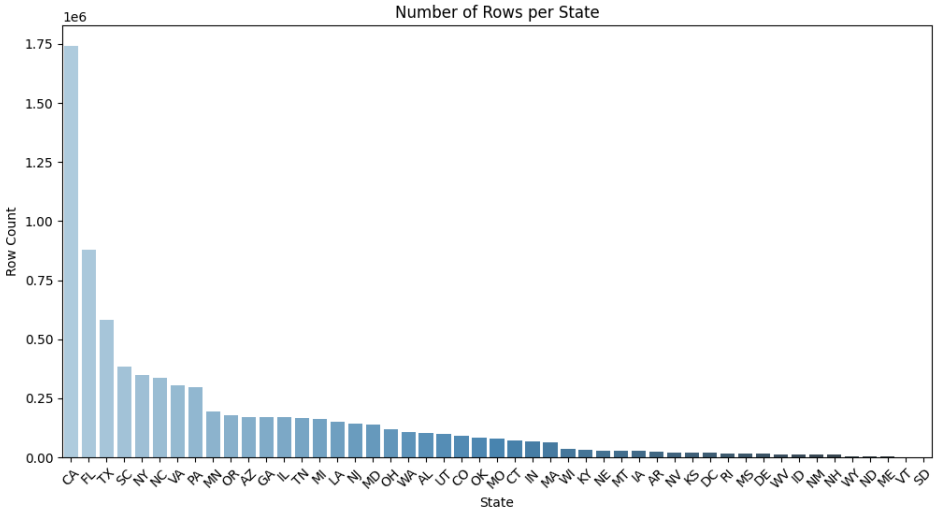


Figure 2. Bar Graph of Number of Accidents by Severity and Week

- Description:
  - It can be seen that an overwhelming number of accidents occur in California. One thing to note is that this is a pure number of cases that does not take into account the number of people living in the state.

## 2.2 Variable 5 - Boolean type column

- Number of True & False:
  - | Name | True | False |
    |---|---|---|
    | Amenity | 96,334 | 7,632,060 |
    | Bump | 3,514 | 7,724,880 |
    | Crossing | 873,763 | 6,854,631 |
    | Give_Way | 36,582 | 7,691,812 |
    | Junction | 571,342 | 7,157,052 |
    | No_Exit | 7,708,849 | 7,708,849 |
    | Railway | 66,979 | 7,661,415 |
    | Roundabout | 249 | 7,728,145 |
    | Station | 201,901 | 7,526,493 |
    | Stop | 214,371 | 7,514,023 |
    | Traffic_Calming | 7,598 | 7,720,796 |
    | Traffic_Signal | 1,143,772 | 6,584,622 |
    | Turning_Loop | 0 | 7,728,394 |

Table 3. Number of True and False Per Boolean Column

- Description:

- Other than No_Exit, the remaining columns have an overwhelming number of false information. There is a limitation that the sample is too small to determine how variables affect thoughts that are true.

# 3. Multivariate analysis

## 3.1 Correlation between Recovery_Time and Severity

- Definition:
  - Recovery_Time: Derived Variable from Variable 2
- Visualiztion:

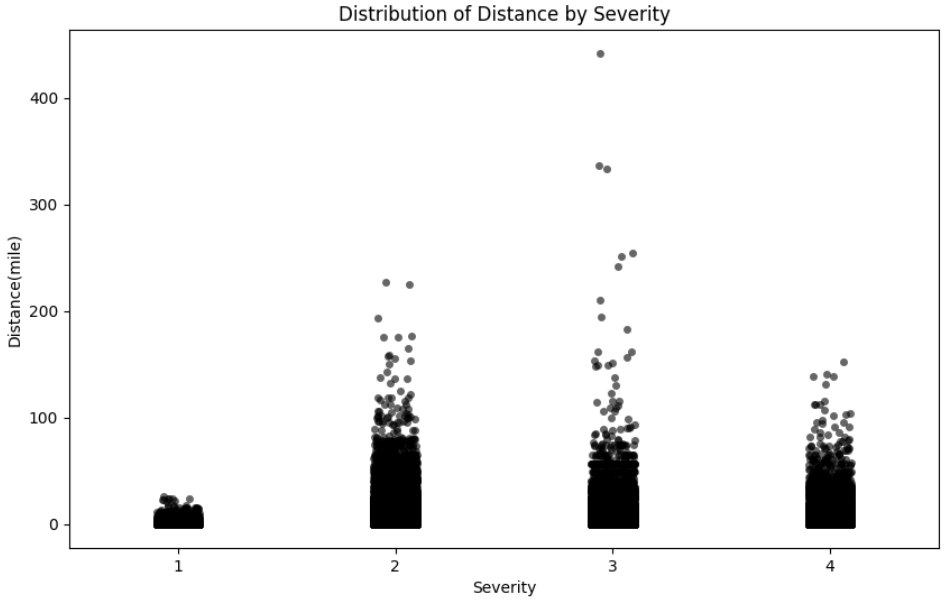


Figure 3. Strip Plot of Affected Distance by Severity

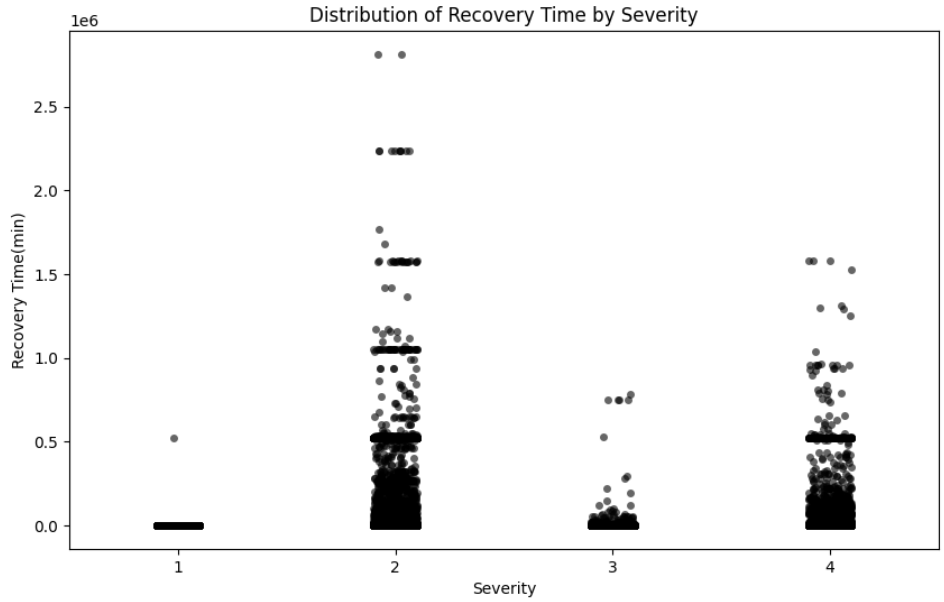


Figure 4. Strip Plot of Revoery Time by Severity

- Description:
  - Severity level 3 causes the second-longest periods of congestion, and has the most extended overall impact distances. Due to this, accidents at severity level 3 are resolved almost as quickly as those at severity level 1. Severity level 2, which congestion distance after level 3, and even most occurring in frequency shows slower recovery times than even severity level 4. This indicates that because accidents are managed based on severity alone, severity 2 accidents—which frequently lead to congestion—are not given priority and tend to be addressed later.

### 3.2 Severity2 Occurrence Frequency by State

- Visualization:

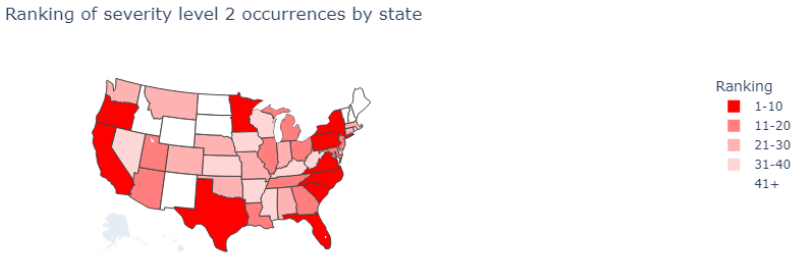


Figure 5. Ranking of severity level 2 occurrences by state

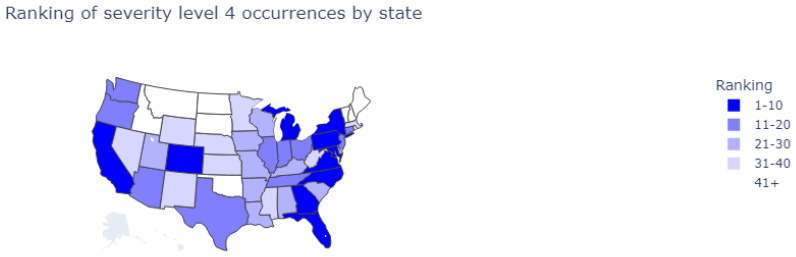


Figure 6. Ranking of severity level 4 occurrences by state

- Description:
  - Let's focus on Severity 2. The red map above visualizes the ranking of areas where Severity 2 incidents occurred most frequently, divided into four units. The areas marked in the darkest red are predominantly large cities, such as those in California, Texas, and New York. On the other hand, the blue map showing the frequency of severity 4 accidents suggests that severity 4 accidents occur more often in high-traffic areas than in large cities. This is supported by the fact that the color of Colorado has become darker and that of Texas and Minnesota has become lighter. In this context, it can be interpreted that Severity 2 accidents primarily occur in urban areas.

## 4. Suggestion

- Severity 2 accidents are the most frequent and are characterized by affecting the longest road distances after Severity 3. However, due to being of a lower severity, the response time for these incidents is

slower than that for severity levels 3 and 4. Notably, since Severity 2 accidents occur more frequently in urban areas compared to Severity 4 accidents, it is essential to prioritize the swift handling of Severity 2 incidents in urban environments.

- A project that could be proposed based on this is to identify the differences between Severity 2 accidents in urban areas and those at other severity levels, in order to prepare for potential accidents in advance or to develop quick response measures. For instance, if it is found that Severity 2 accidents occur more frequently around airports compared to other severity levels, it is important to understand the psychology of drivers who are in a hurry to make their takeoff and landing times. This understanding could lead to actions such as installing additional roadblocks or establishing multiple accident response teams around the airport. Through these measures, we can expect to minimize unnecessary urban congestion as well as material and human damage. While airport and collision data may contain many NA values, there is potential for improvement by combining them with additional data or using predicted values generated through a prediction model.