Exploratory Data Analysis

- StudentID: 22100784

- Name: 한희

- 1st Major: 상담심리

- 2nd Major: 데이터사이언스

주별 사고 발생 빈도를 분석하고, 기상 조건, 도로 상태, 기타 요인과의 상관관계를 통해 사고 원인을 파악하여 해결방안을 탐구해보고자 한다.

1. Data overview

- Sample size

```
rows:  7728394
columns:  46
```

- Variables

```
- Start_Time(datetime64): 사고 시작 시간
- Start_Lat, Start_Lng(float64): 사고 시작 지점의 위도와 경도
- State(string): 사고가 발생한 주
- Visibility(mi)(float64): 사고 발생 당시의 가시거리
- Weather_Condition(float64): 사고 당시의 날씨 상태
- Month, Hour(int32): 사고가 발생한 월과 시간
- Bump(bool): 사고 지역 근처에 과속 방지턱 또는 과속 방지 요철이 있는지 여부
- Crossing(bool): 사고 지역 근처에 횡단보도가 있는지 여부
- Give_Way(bool): 사고 지역 근처에 양보 표지판이 있는지 여부
- Junction(bool): 사고 지역 근처에 교차로가 있는지 여부
- Turning_Loop(bool): 사고 지역 근처에 회전 구간이 있는지 여부
- Traffic_Signal(bool): 사고 지역 근처에 신호등이 있는지 여부
```

주별 사고 빈도를 파악하고, 그 원인을 다방면으로 분석하기 위해 총 46개의 변수 중 유의미할 것 같은 변수 총 11개를 선택하였다. 또한, 변수 간의 상관 정도를 파악할 수 있도록 'Month', 'Hour' 컬럼을 추가하였으며, 'Weather_Condition' 값을 범주형에서 수치형으로 변형하였다.

2. Univariate analysis

사고 빈도수에 초점을 맞춰 사고 발생 빈도가 높은 지역, 시간대, 날씨 등에 대해 알아보고자 한다.

2.1 주별 사고 분석

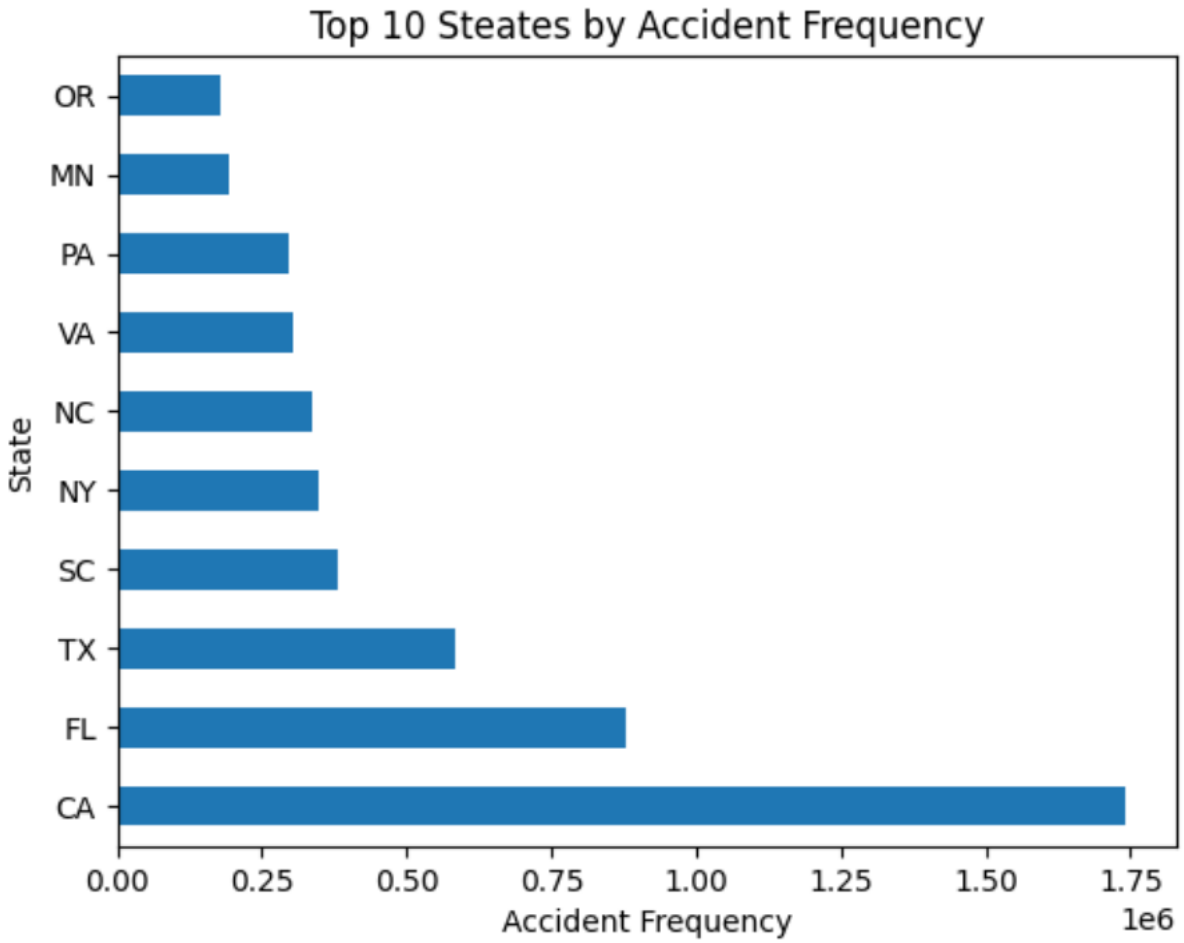

Figure1. 주별 사고 빈도수

위의 그래프를 보면 캘리포니아의 사고 빈도수가 가장 높고, 플로리다와 텍사스가 그 뒤를 따르고 있음을 알 수 있다. 미국 주별 인구수와 비슷한 흐름으로 가고 있는 것으로 보아 인구수가 사고에 영향을 미치고 있음을 알 수 있다.

2.2 날씨별 사고 분석

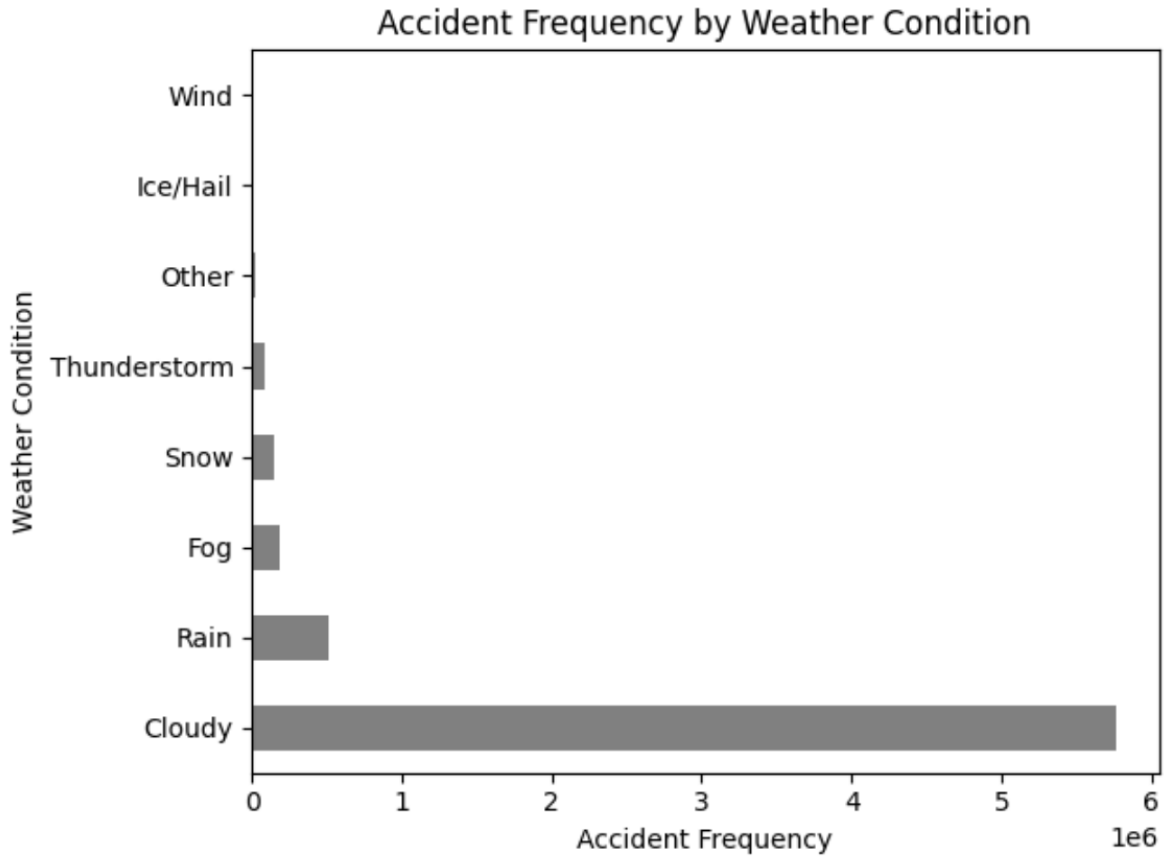

Figure2. 날씨 상태에 따른 사고 빈도수

위의 그래프를 통해 흐림(Cloudy), 비(Rain), 안개(Fog)와 같은 날씨에서 사고가 가장 빈번하게 발생한다는 것을 확인할 수 있다. 이는 날씨 조건이 사고에 중요한 요인으로 작용한다는 것을 시사한다고 볼 수 있으며, 가시거리가 제한되는 조건에서 사고의 빈도수가 높아진다는 패턴을 파악할 수 있다.

2.3 시간대별 사고 분석

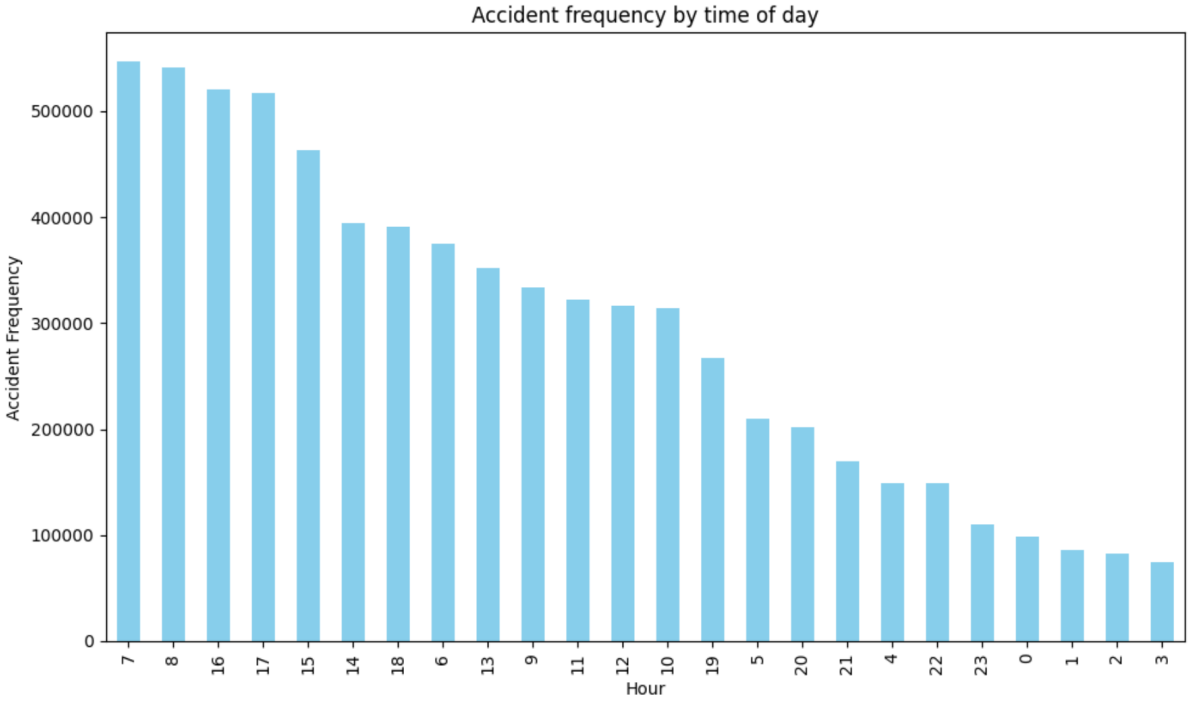

Figure3. 시간대별 사고 빈도수

위의 그래프를 통해 오전 7시 ~ 8시, 오후 4시 ~ 5시에 사고가 빈번하게 일어남을 확인할 수 있다. 이를 통해 차가 많이 몰리는 출퇴근 시간에 사고가 빈번하게 일어난다는 것을 파악할 수 있다.

2.4 사고 발생 요인 분석

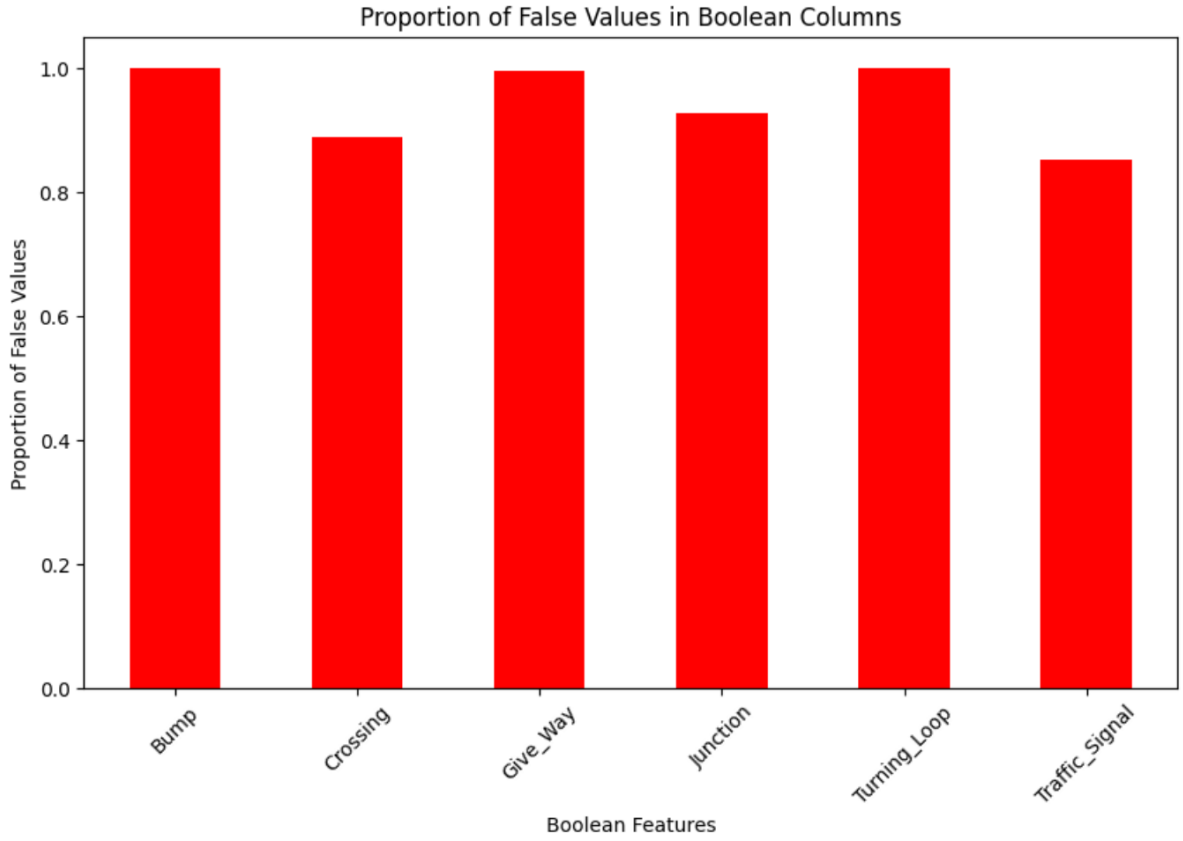

Figure4. 사고 발생 요인별 False 값 비율

위 그래프를 통해 과속 방지턱(Bump), 횡단보도(Crossing), 교차로(Junction), 신호등(Traffic_Signal) 등의 환경 요소가 없는 값이 사고 발생 구간에서 매우 높은 비율을 차지하는 것을 확인할 수 있다. 이는 사고가 발생한 도로의 대부분이 환경 요소가 없는 구간이라는 것을 시사한다고 볼 수 있다.

3. Multivariate analysis

3.1 Correlation

	Visibility(mi)	Weather_Condition	Bump	Crossing	Give_Way	Junction	State_Accident_Frequency
Visibility(mi)	1.000000	0.413727	0.004490	0.038487	0.002704	-0.006197	0.022014
Weather_Condition	0.413727	1.000000	0.001471	0.022392	0.001604	-0.013125	0.046986
Bump	0.004490	0.001471	1.000000	0.012070	0.000386	-0.004333	0.002497
Crossing	0.038487	0.022392	0.012070	1.000000	0.057997	-0.087737	-0.060854
Give_Way	0.002704	0.001604	0.000386	0.057997	1.000000	-0.009283	-0.027457
Junction	-0.006197	-0.013125	-0.004333	-0.087737	-0.009283	1.000000	0.049756
State_Accident_Frequency	0.022014	0.046986	0.002497	-0.060854	-0.027457	0.049756	1.000000

Figure5. 주요 변수들 간의 상관관계

위의 그래프를 보면 가시거리(Visibility)와 날씨 상태(Weather_Condition) 간에는 어느 정도의 양의 상관관계가 있는 것으로 보아 날씨 상태가 안 좋을수록 가시거리가 감소한다는 경향을 파악할 수 있다. 또한, 주별 사고 빈도는 대부분의 변수들과 약한 상관관계를 보임을 확인할 수 있다. 따라서 환경적 요인들이 사고 빈도에 어떤 영향을 미치는지 파악하는 데 도움이 될 수 있지만, 변수들 간의 관계는 매우 약함을 알 수 있다.

4. Suggestion

1. 인구밀도가 높은 지역

Problem: 인구밀도가 높은 지역에서 사고 빈도가 높다는 것을 알 수 있다. 이는 인구가 많을수록 교통량이 증가하고, 사고 위험이 더 커질 수 있음을 시사한다.
Solve: 대중교통의 이용을 장려하여 도로에서 개인 차량의 비율을 줄임으로 교통량을 줄이는 방법있다. 또한, 한 Street의 끝에서 일정한 속도로 운전하게 되면 연속적인 통과 신호를 부여 받아 멈추지 않고 통행하는 새로운 신호 체계를 도입함으로써 교통량을 실시간으로 조절하여 통행 시간을 효과적으로 줄이는 방법을 탐구해볼 수 있다.

2. 날씨 조건에 따른 안전

Problem: 흐림(Cloudy), 비(Rain), 안개(Fog)와 같은 날씨 조건에서 사고가 빈번하게 발생하는 경향이 있으며, 가시거리가 제한되는 조건에서 사고 빈도가 높아진다.
Solve: 날씨가 좋지 않은 날에는 스마트 도로 경고 시스템을 통해 운전자에게 속도를 줄이도록 알리는 방법을 활용할 수 있다. 또한, 가시거리가 제한되는 구간에서는 추가적인 조명을 설치하거나 반사형 도로 표지판 등을 추가로 설치하는 방법을 활용할 수 있다.