Exploratory Data Analysis

- StudentID: 22101059

- Name: 이예정

- 1st Major: 도시환경공학

- 2nd Major: 데이터사이언스

1. 미국에서 사고가 빈번하게 발생하는 주를 발견하고 그 주의 특징을 파악한다.
2. 교통 사고에 날씨 변수들이 어떠한 영향을 미치는지 분석한다.

1. Data overview

1-1. Data informaiton

- Rows: 7728394 rows
- Columns: 46 columns
- dtypes: bool(13), float64(12), int64(1), object(20)

1-2. Data describe

	Severity	Start_Lat	Start_Lng	End_Lat	End_Lng	Distance(mi)	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
count	7728394.00	7728394.00	7728394.00	4325632.00	4325632.00	7728394.00	7564541.00	5729375.00	7554250.00	7587715.00	7551296.00	7157161.00	5524808.00
mean	2.21	36.20	-94.70	36.26	-95.73	0.56	61.66	58.25	64.83	29.54	9.09	7.69	0.01
std	0.49	5.08	17.39	5.27	18.11	1.78	19.01	22.39	22.82	1.01	2.69	5.42	0.11
min	1.00	24.55	-124.62	24.57	-124.55	0.00	-89.00	-89.00	1.00	0.00	0.00	0.00	0.00
25%	2.00	33.40	-117.22	33.46	-117.75	0.00	49.00	43.00	48.00	29.37	10.00	4.60	0.00
50%	2.00	35.82	-87.77	36.18	-88.03	0.03	64.00	62.00	67.00	29.86	10.00	7.00	0.00
75%	2.00	40.08	-80.35	40.18	-80.25	0.46	76.00	75.00	84.00	30.03	10.00	10.40	0.00
max	4.00	49.00	-67.11	49.08	-67.11	441.75	207.00	207.00	100.00	58.63	140.00	1087.00	36.47

- Severity: 평균이 2.21로 전체적으로 사고 심각도가 높지 않다.
- Distance(mi): 최대값이 442mi로 평균에 비해 너무 높아 outlier인듯하다. (442mi은 서울에서 부산까지 거리의 두 배정도에 해당)
- Temperature(F), Wind_Chill(F): 최대값이 207F로 너무 높아 outlier인듯하다. (207F = 97°C)
- Visibility(mi): 최대값이 140mi로 너무 높아 outlier인듯히다.
- Wind_Speed(mph): 최대값이 1087mph로 너무 높아 outlier인듯하다.
- Precipitation(in): 최대값이 37in로 너무 높아 outlier인듯하다.

1-3. Variables

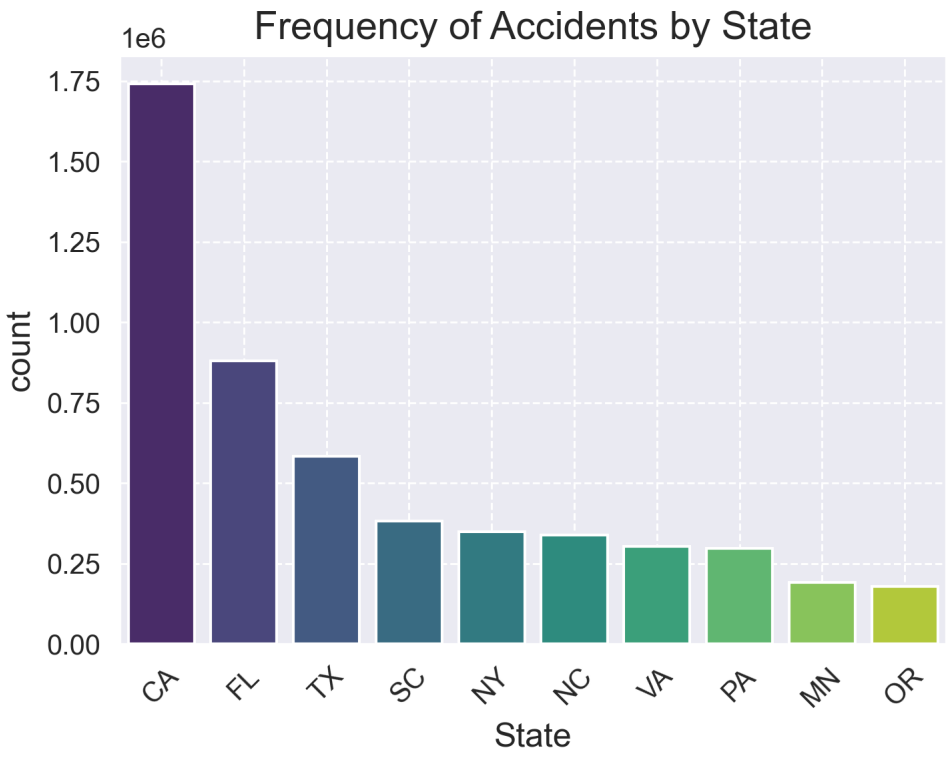
해당 데이터의 변수는 크게 4가지로 구분할 수 있다.

- 사고 기본 정보: ID, 사고심각도, 발생 시간, 지역이름 등 사고 발생 관련 기본 정보를 알 수 있는 변수들
- 날씨 정보: 사고일의 온도, 습도, 강수량 등 기상 정보를 알 수 있는 변수들
- POI(관심지점) 정보: 사고 발생지 근처에 교통 관련 시설물, 신호등, 횡단보도 등이 있는지 여부를 알 수 있는 변수들

- 시간대 정보: 시간적으로 언제 발생했는지 알 수 있는 변수들

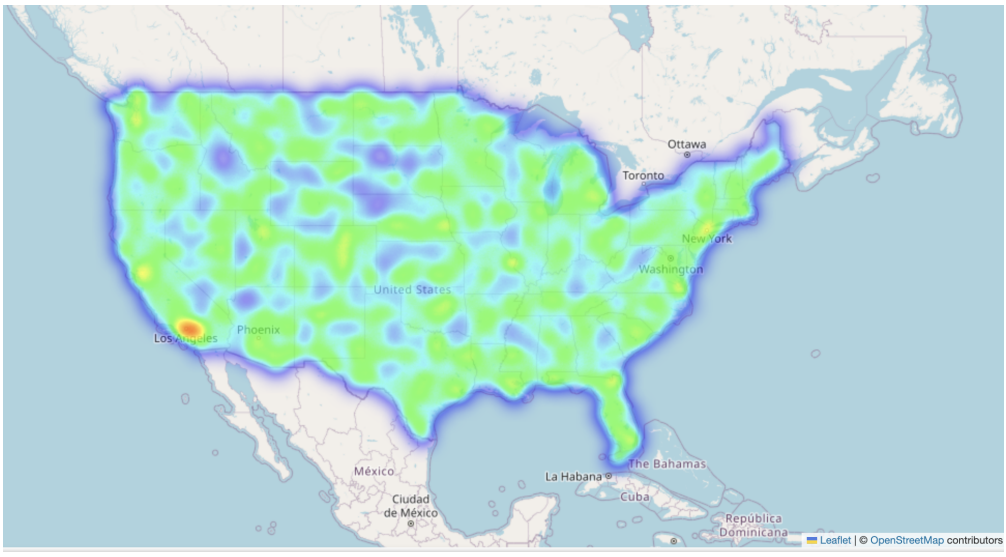
2. Univariate analysis

2-1. 사고 빈도 비교 by State

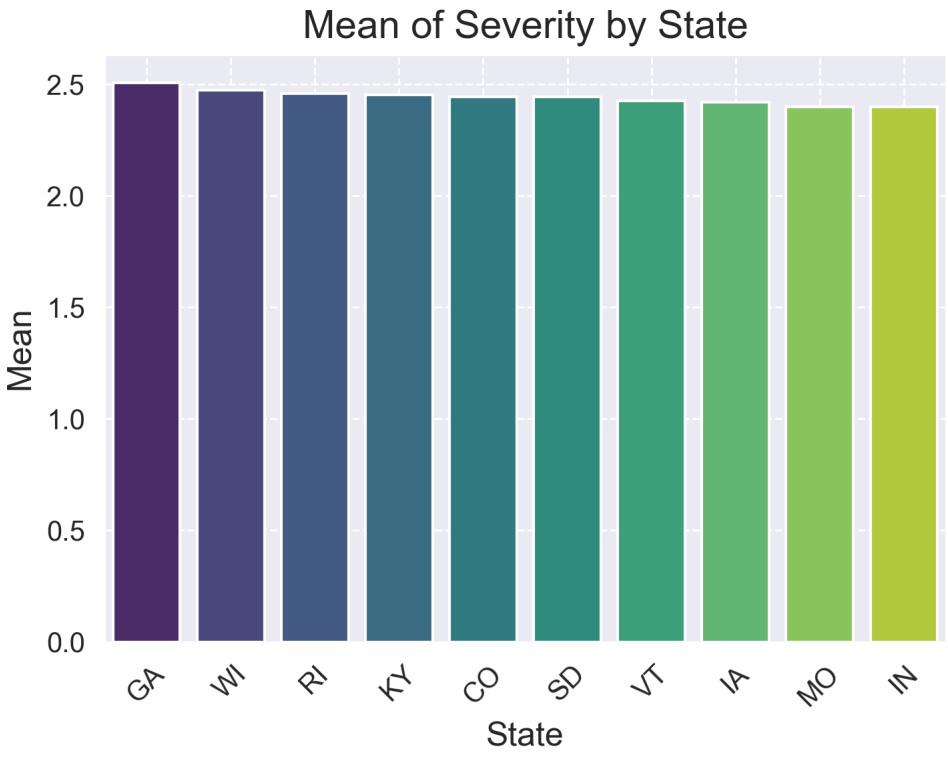


각 주의 사고 빈도를 계산하였다. 상위 10개 주만 나타내었고 가장 높은 주는 CA, 즉 California(캘리포니아)이다. 그 다음으로는 FL(플로리다), TX(텍사스) 각각 2, 3순위를 차지한다. 하지만 캘리포니아의 경우, 2순위인 플로리다와 2배 가량 차이가 나며 사고가 가장 빈번한 주로 뚜렷하게 나타난다. 캘리포니아의 높은 사고 빈도의 배경 및 원인에 대해 면밀히 살펴볼 필요가 있다고 생각한다.

*아래 지도를 보면 가장 빨간 부분이 캘리포니아다.

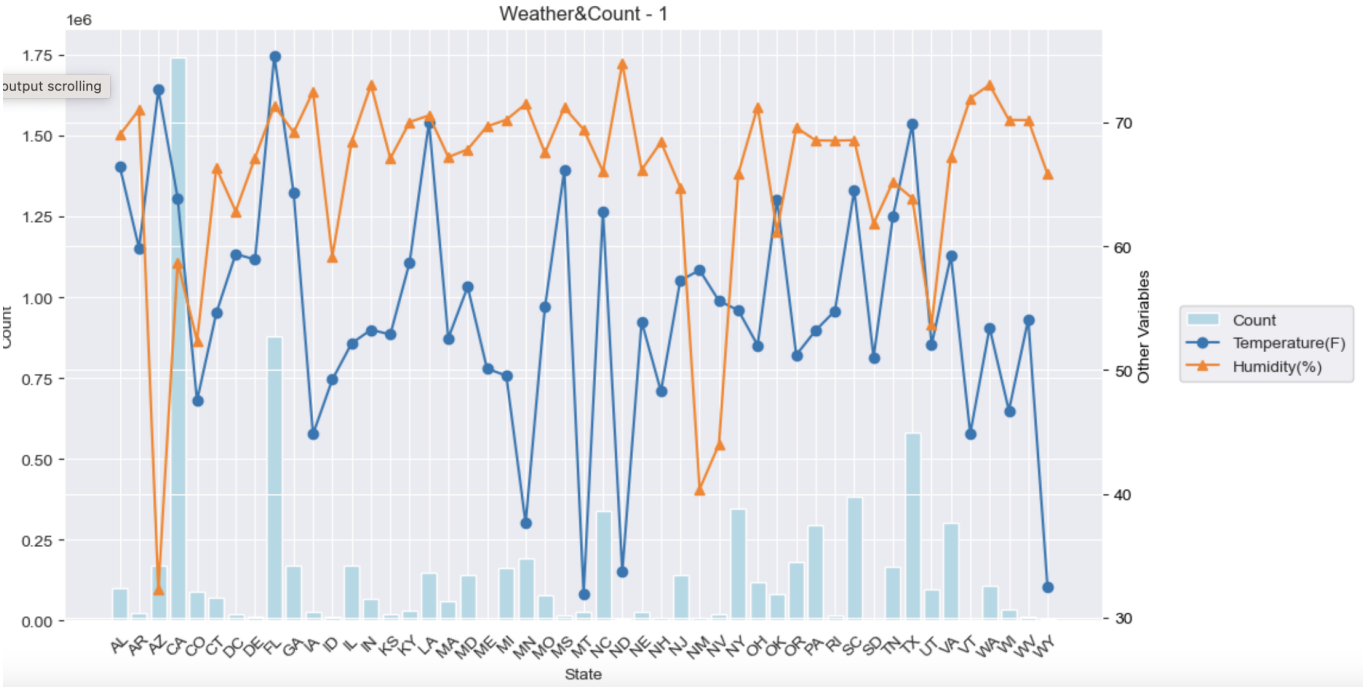


2-2. 사고 심각도 평균 비교 by State

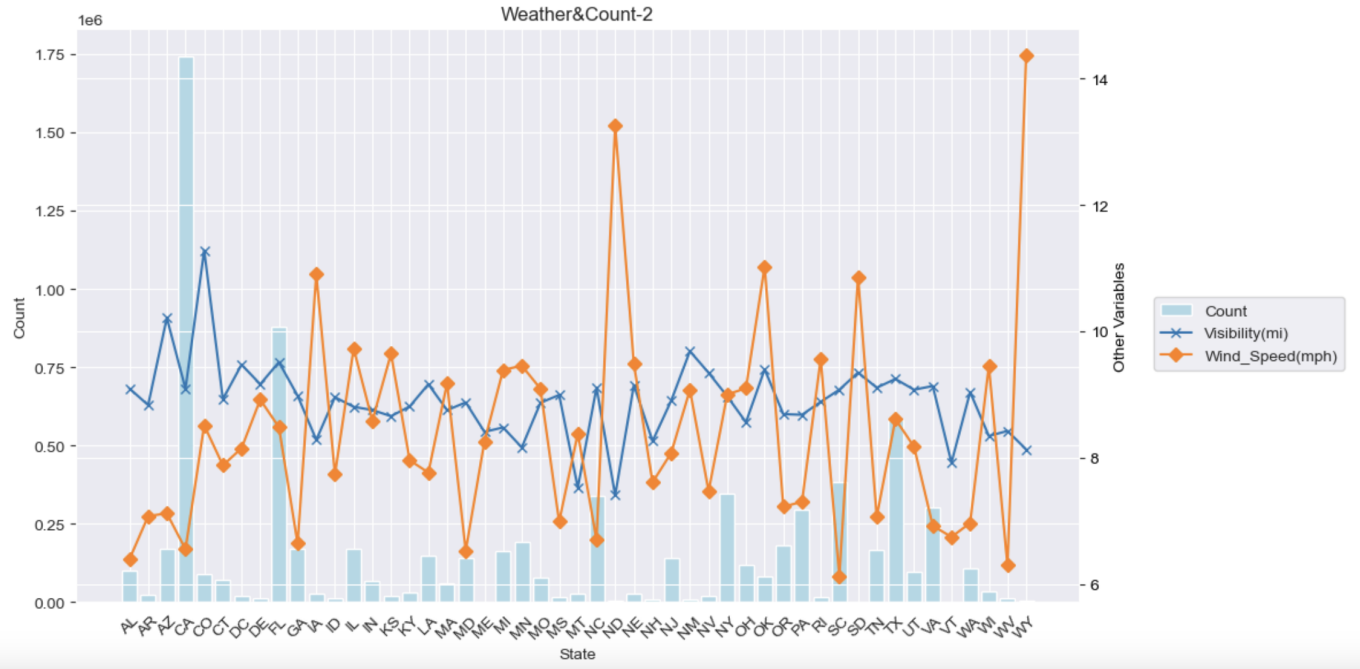


- 사고 심각성은 GA(조지아), WI(위스콘신), RI(로드아일랜드) 순으로 높다.
- 가장 높은 점수가 4점임을 감안할때 사고 심각도가 뚜렷하게 높은 주는 없음을 알 수 있다.
- 사고 빈도가 가장 높았던 캘리포니아의 경우 사고 심각도의 평균은 2.17로 전체 평균(2.27)보다 낮았다.

2-3. 날씨정보와 사고 빈도수 비교



- 위 그래프는 주별로 사고 빈도수와 온도, 습도를 나타낸 그래프다.
- 사고 빈도수가 가장 높은 캘리포니아, 플로리다, 텍사스의 경우 온도와 습도가 다소 높은 편이다.



- 위 그래프는 주별로 사고 빈도수와 시정, 풍속을 나타낸 그래프다.
- 사고 빈도수가 가장 높은 캘리포니아, 플로리다, 텍사스의 경우 풍속이 다소 낮은 편이고, 시정은 평균 정도이다.

3. Multivariate analysis

3-1. 상관관계 분석

	count	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)
count	1.000000	0.399744	0.420967	-0.058140	0.200737	0.177772	-0.249712	-0.052042
Temperature(F)	0.399744	1.000000	0.980424	-0.243871	0.484270	0.614539	-0.542542	0.297980
Wind_Chill(F)	0.420967	0.980424	1.000000	-0.241445	0.430397	0.598620	-0.592773	0.246331
Humidity(%)	-0.058140	-0.243871	-0.241445	1.000000	0.492490	-0.631826	0.078106	0.435160
Pressure(in)	0.200737	0.484270	0.430397	0.492490	1.000000	-0.084168	-0.373557	0.594035
Visibility(mi)	0.177772	0.614539	0.598620	-0.631826	-0.084168	1.000000	-0.269045	-0.146924
Wind_Speed(mph)	-0.249712	-0.542542	-0.592773	0.078106	-0.373557	-0.269045	1.000000	-0.267150
Precipitation(in)	-0.052042	0.297980	0.246331	0.435160	0.594035	-0.146924	-0.267150	1.000000

미국의 주별로 사고 빈도수를 계산하고 사고 빈도수와 날씨 요인 간에 어떤 상관관계가 있는지 파악하였다.

- 양(+): 온도, 체감온도, 기압, 시정
- 음(-): 습도, 풍속, 강수량

온도, 체감온도: 온도가 높을수록 사고가 더 발생한다. 사고 빈도수가 가장 높았던 캘리포니아는 지중해성 기후, 플로리다는 열대 기후, 텍사스는 아열대 기후로 대체로 따뜻한 지역들이다.
기압: 기압이 높을수록 사고 빈도가 더 높다. CA, FL, TX의 경우 전체 기압 평균보다 높다.
시정: 시정이 클수록 사고 빈도가 더 높다. CA, FL, TX의 경우 전체 시정 평균보다 높다.
습도: 습도가 낮을수록 사고 빈도가 더 높다. 하지만 상관계수가 0.05로 매우 낮기에 거의 영향을 미치지 않는다.
풍속: 속도가 낮을수록 사고 빈도가 더 높다. CA는 속도가 평균보다 낮고 FL, TX의 경우 평균보다 높다.
강수량: 강수량이 낮을수록 사고 빈도가 더 높다. 하지만 상관계수가 0.05로 매우 낮기에 거의 영향을 미치지 않는다.

4. Suggestion

1. 사고 빈도수가 높았던 주들의 기후적 특징 고려
    - 캘리포니아, 플로리다 등 사고 빈도수가 높았던 주들의 기후적 특징이 비슷하게 나타났다. 따라서 사고 빈도와 기후와의 연관성을 고려하여 날씨에 따른 사고 예방 대책을 세울 수 있다.
    - 주별 사고 빈도수와 날씨 정보 변수들의 회귀분석을 통해 어떠한 변수가 어떻게 영향을 유의미하게 미치는지 알아볼 수 있다.
2. 추가로 더 분석하고 싶은 점
    - 사고 빈도수가 높은 지역들을 중심으로 POI 분포를 파악할 필요가 있다고 생각한다.
    - 특히, 사고를 예방할 수 있는 교통 시설의 파악을 통해 이러한 시설의 부족으로 인한 것은 아닌지 알아볼 필요가 있다.