Exploratory Data Analysis

- StudentID: 21900471

- Name: 육정인

- 1st Major: 경제학

- 2nd Major: 데이터 사이언스

이 EDA 보고서는 자동차 사고에 대한 데이터를 세가지 측면 (1) 시간적 측면 (2) 기후적 측면 (3) 도로환경적 측면에서 접근할 것이다.
단변량분석과 다변량분석을 통해 어떠한 변수들이 사고의 발생빈도와 심각도에 영향을 미치는지 분석할 것이다.
이를 통해 사고를 예방하는데 도움을 줄 수 있을 것으로 기대된다.

1. Data overview

- Data: accidents_data.csv
- Description: 자동차 사고와 관련된 심각도, 시간, 위치, 기후, 주변환경에 대한 정보가 담겨있다.
- Sample Size: 7728394x46
- Variables:
    1. ID(object): 해당사고에 대한 고유번호
    2. Source(object): 사고 데이터의 출처
    3. Severity(int64): 심각도 (1부터 4까지 있으며 낮을수록 덜 심각한 사고이고 높을수록 더 심각한 사고이다)
    4. Start_Time(object): 사고가 일어난 시간
    5. End_Time(object): 사고가 끝난 시간
    6. Start_Lat(float64): 사고가 시작된 위치의 위도
    7. Start_Lng(float64): 사고가 시작된 위치의 경도
    8. End_Lat(float64): 사고가 종료된 위치의 위도
    9. End_Lng(float64): 사고가 종료된 위치의 경도
    10. Distance(mi)(float64): 사고가 발생한 도로의 길이
    11. Description(object): 사고에 대한 설명
    12. Street(object): 도로명
    13. City(object): 도시
    14. County(object): 군 (주보다 하위개념)
    15. State(object): 주
    16. Zipcode(object): 우편번호
    17. Country(object): 나라
    18. Timezone(object): 시간대
    19. Airport_Code(object): 사고지점과 가장 가까운 공항
    20. Weather_Timestamp(object): 기후를 관측한 시점
    21. Temperature(F)(float64): 온도(화씨)

22. Wind_Chill(F)(float64): 체감온도(화씨)
23. Humidity(%)(float64): 습도
24. Pressure(in)(float64): 기압
25. Visibility(mi)(float64): 가시거리
26. Wind_Direction(object): 풍향
27. Wind_Speed(mph)(float64): 풍속
28. Precipitation(in)(float64): 강수량
29. Weather_Condition(object): 날씨상태
30. Amenity(bool): 주변 편의시설 여부
31. Bump(bool): 방지턱 여부
32. Crossing(bool): 횡단보도 여부
33. Give_Way(bool): 양보 표지판 여부
34. Junction(bool): 교차로 여부
35. No_Exit(bool): 출구없음 표지판 여부
36. Railway(bool): 철도길 여부
37. Roundabout(bool): 회전 교차로 여부
38. Station(bool): 역 또는 정류장 여부
39. Stop(bool): 정지 표지판 여부
40. Traffic_Calming(bool): 교통정온화 여부
41. Traffic_Signal(bool): 신호등 여부
42. Turning_Loop(bool): 유턴 가능 구역 여부
43. Sunrise_Sunset(object): 일몰과 일출을 기준으로 낮이였는지 밤이였는지 구분
44. Civil_Twilight(object): 시민박명 구분(태양이 수평선 아래 0도에서 6도 사이에 있을 때, 조명없이 야외활동이 가능하며 글씨를 읽을 수 있는 정도)
45. Nautical_Twilight(object): 항해박명 구분(태양이 수평선 아래 6도에서 12도 사이에 있을 때, 조명없이 야외활동은 힘들지만 사물 구분이 가능한 정도)
46. Astronomical_Twilight(object): 천문박명 구분(태양이 수평선 아래 12도에서 18도 사이에 있을 때, 조명없이 야외활동이 불가능하며 별을 관측할 수 있다)

2. Univariate analysis

2.0 Severity

사고의 심각도를 파악하기 위해서 Severity 변수를 선택하였다. Severity변수는 범주형 변수로 1부터 4까지의 값을 가진다. 1은 가장 경미한 사고, 4는 가장 심각한 사고임을 의미한다.

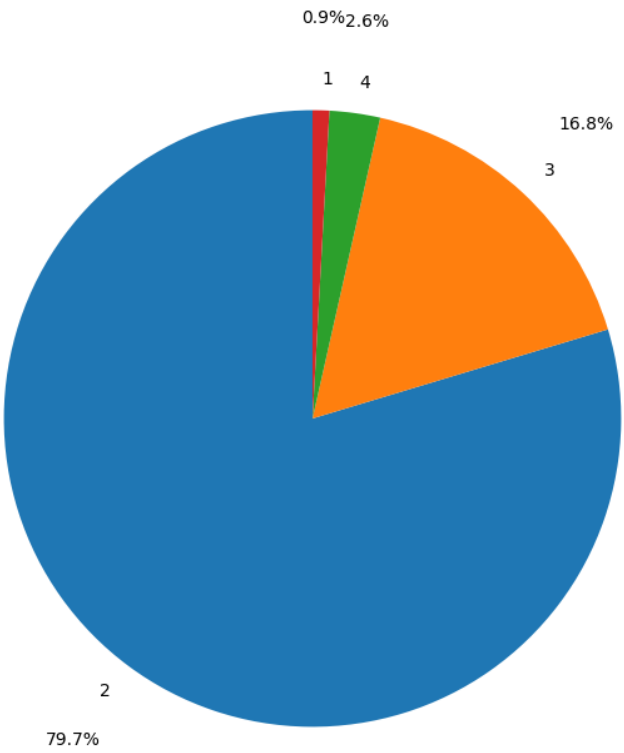

Figure1. Severity Distribution

표1를 통해 사고의 심각도는 2가 가장 많은 비중을 차지하고 있으며 그다음 3,4,1 순으로 많은 비중을 차지하고 있음을 알 수 있다.
약 80% 가량의 사고가 가벼운 사고였음을 알 수 있다.

2.1 시간적 측면

사고를 시간적 측면에서 접근하기 위해 선택한 변수는 다음과 같다.
'Start_Time', 'Sunrise_Sunset', 'Civil_Twilight', 'Nautical_Twilight', 'Astronomical_Twilight'

먼저 'Start_Time'변수를 선택하여 사고가 일어난 시점을 기준으로 분석하였고 년도, 월, 시간대, 요일별로 그룹화하여 사고의 빈도수를 분석하였다.

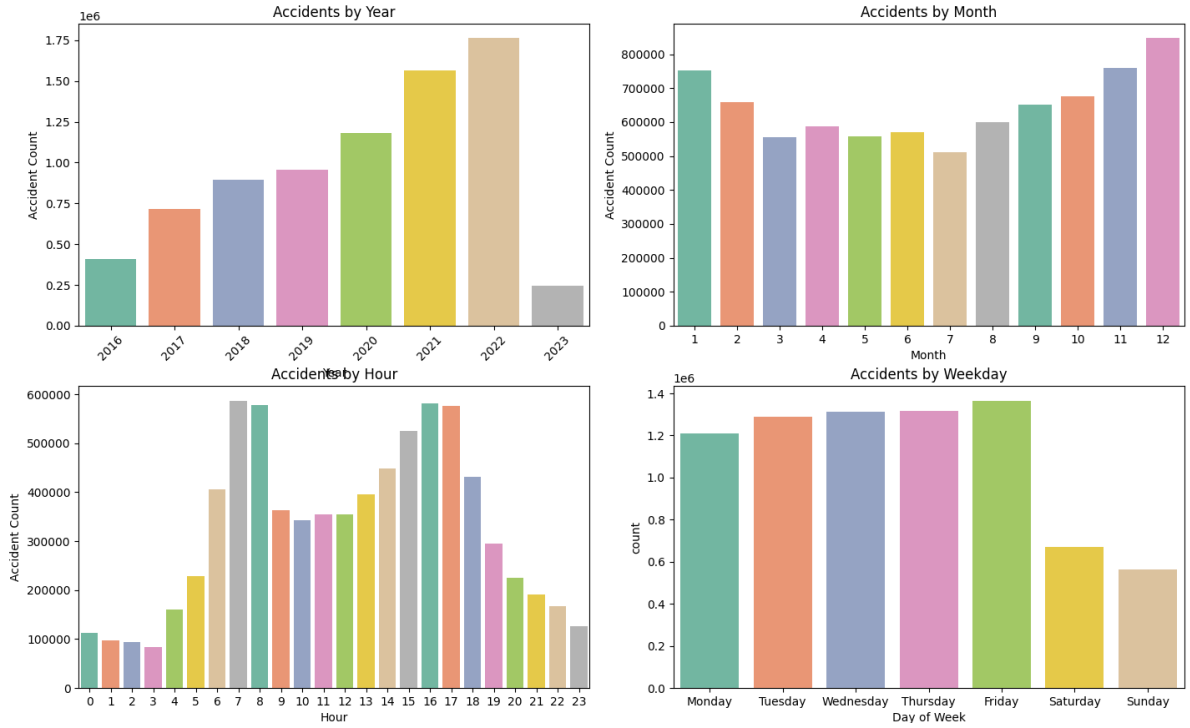

Figure2. Accident Count by Year, Month, Hour, Weekday

표2에 년도는 먼저 보면 2016년부터해서 2022년까지 사고가 증각하다가 2023년에 급감한 것을 볼 수 있다. 하지만 2023년 데이터는 3월 31일까지만 기록되어 있기 때문에 데이터의 표본이 다른 년도보다 적음을 감안해야 한다. 월별 추이를 보면 11월부터 1월까지 겨울시즌에 가장 사고가 많음을 알 수 있다. 시간대를 보면 7시부터 8시까지 그리고 16시부터 17시까지 가장 사고가 많음을 알 수 있는데 이는 출퇴근 시간에 사고가 많이 발생하는 것으로 예상해 볼 수 있다. 요일별 추이를 보면 주말보다는 평일에, 특히 금요일에 가장 사고가 많이 발생하는 것을 볼 수 있다.
만약 경찰이라면 11월부터 1월사이에 그리고 평일 출퇴근 시간대에 감시인력을 더 늘려 사고를 줄일 수 있을 것이다.

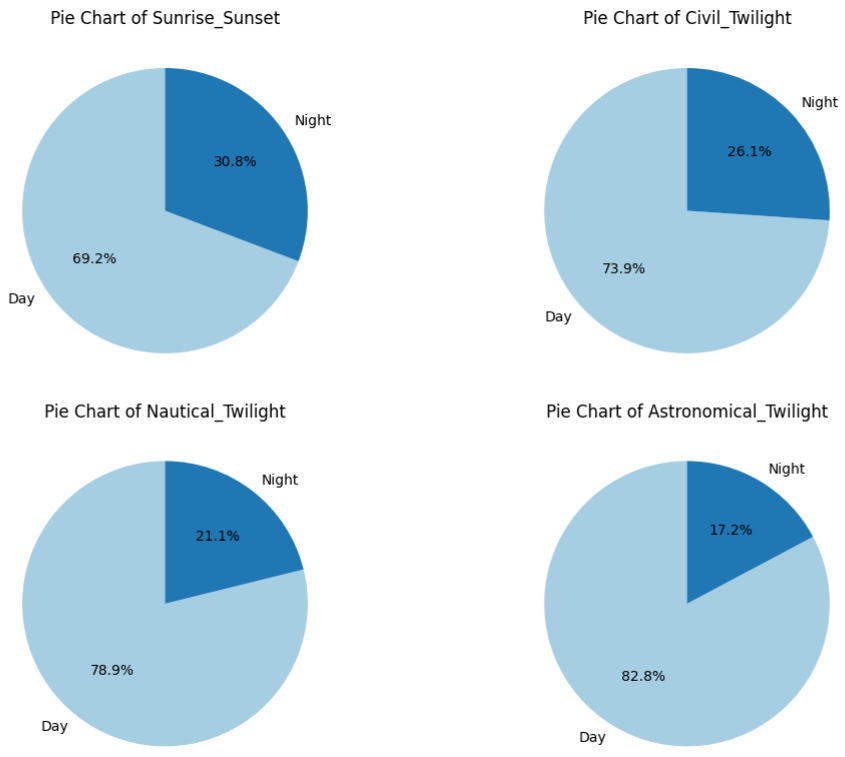

Figure3. Distribution of Day and Night at Various Twhilight

위의 파이차트는 사고가 일어났을 당시 해가 어느 위치에 있었는지를 알려준다. Sunrise_Sunset(일몰일출), Civil_Twilight(시민박명), Nautical_Twilight(항해박명), Astronomical_Twilight(천문박명)은 각각 해가 수평선 아래 0도, 6도, 12도, 18도보다 위에 있는지 아래에 있는지 알려준다. 해는 수평선 아래로 내려가도 어느정도 빛이 남아있기 때문에 해의 위치에 따라 야외에서 시야식별여부가 달라진다. 표3을 보면 해가 수평선 아래로 내려갈수록 밤에 발생하는 사고의 비율이 점점 줄어든는 것을 확인할 수 있다. 해가 수평선 아래로 내려가 더 어두워질수록 사고가 더 많이 발생할 것으로 예상했지만 반대로 더 줄어드는 것을 알 수 있었다. 하지만 이는 밤에 운전하는 사람들이 줄어드는 것일 수도 있기 때문에 해당 데이터만으로는 판단하기 어렵다.

2.2 기후적 측면

사고 데이터를 기후적 측면에서 접근하기 위해 선택한 기후 관련 변수는 다음과 같다.
'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', 'Pressure(in)', 'Visibility(mi)', 'Wind_Speed(mph)', 'Precipitation(in)'
해당 변수들은 모두 연속형 변수들이다. 해당 변수들의 분포를 파악하기 위해 히스토그램을 이용해 시각화했으며 KDE옵션을 사용해 확률밀도도 같이 시각화했다. 결측치는 모두 제외하고 계산하였다.

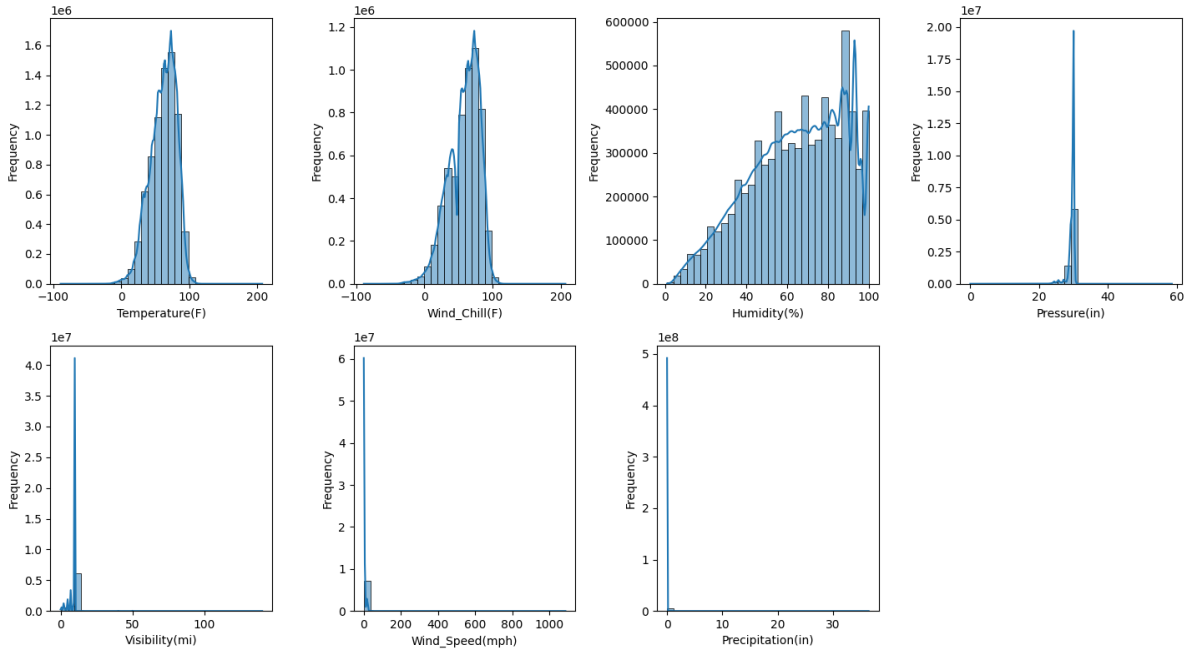

Figure4. Histogram with KDE of Weather Data

표4를 보면 온도(Temperature)와 체감온도(Wind_Chill)은 상당히 비슷한 분포를 보이고 있다. 둘다 화씨60에서 화씨80에 가장 많은 분포를 이루고 있으며 이는 1년 중 가장 많이 관측되는 기온이기 때문에 그럴 것으로 예상된다. 기압(Pressure), 가시거리(Visibility), 풍속(Wind_Speed), 강수량(Precipitation)은 특정 값에 거의 모든 분포가 집중되어 있음을 볼 수 있다. 습도(Humidity)만이 특정한 패턴을 보이고 있으며 습도가 높아짐에 따라 사고 빈도도 높아지는 것을 확인할 수 있다. 이는 습도가 높을 때 비나 안개, 눈이 오기 때문으로 예상된다.

날씨상태(Weather Condition)변수는 범주형 데이터로 맑음, 구름낌, 눈, 등등의 상태로 표현된다. 날씨상태 변수는 고유값으로 144개의 값을 가지고 있다. 하지만 설명적인 특징으로 인해 겹치는 고유값이 많아 가장 많이 나타나는 Rain과 Cloudy를 포함하고 있는 값은 모두 하나로 합쳤다. 그리고 100,000개 미만으로 나타나는 값들은 Other로 합쳤다.

Unique values of Weather_Condition: 'Light Rain', 'Overcast', 'Mostly Cloudy', 'Rain', 'Light Snow', 'Haze',

'Scattered Clouds', 'Partly Cloudy', 'Clear', 'Snow', 'Light Freezing Drizzle', 'Light Drizzle', 'Fog', 'Shallow Fog', 'Heavy Rain', 'Light Freezing Rain', 'Cloudy', 'Drizzle', nan, 'Light Rain Showers', 'Mist', 'Smoke', 'Patches of Fog', 'Light Freezing Fog', 'Light Haze', 'Light Thunderstorms and Rain', 'Thunderstorms and Rain', 'Fair', 'Volcanic Ash', 'Blowing Sand', 'Blowing Dust / Windy', 'Widespread Dust', 'Fair / Windy', 'Rain Showers', 'Mostly Cloudy / Windy', 'Light Rain / Windy', 'Hail', 'Heavy Drizzle', 'Showers in the Vicinity', 'Thunderstorm', 'Light Rain Shower', 'Light Rain with Thunder', 'Partly Cloudy / Windy', 'Thunder in the Vicinity', 'T-Storm', 'Heavy Thunderstorms and Rain', 'Thunder', 'Heavy T-Storm', 'Funnel Cloud', 'Heavy T-Storm / Windy', 'Blowing Snow', 'Light Thunderstorms and Snow', 'Heavy Snow', 'Low Drifting Snow', 'Light Ice Pellets', 'Ice Pellets', 'Squalls', 'N/A Precipitation', 'Cloudy / Windy', 'Light Fog', 'Sand', 'Snow Grains', 'Snow Showers', 'Heavy Thunderstorms and Snow', 'Rain / Windy', 'Heavy Rain / Windy', 'Heavy Ice Pellets', 'Light Snow / Windy', 'Heavy Freezing Rain', 'Small Hail', 'Heavy Rain Showers', 'Thunder / Windy', 'Drizzle and Fog', 'T-Storm / Windy', 'Blowing Dust', 'Smoke / Windy', 'Haze / Windy', 'Tornado', 'Light Drizzle / Windy', 'Widespread Dust / Windy', 'Wintry Mix', 'Wintry Mix / Windy', 'Light Snow with Thunder', 'Fog / Windy', 'Snow and Thunder', 'Light Snow Shower', 'Sleet', 'Light Snow and Sleet', 'Snow / Windy', 'Rain Shower', 'Snow and Sleet', 'Light Sleet', 'Heavy Snow / Windy', 'Freezing Drizzle', 'Light Freezing Rain / Windy', 'Thunder / Wintry Mix', 'Blowing Snow / Windy', 'Freezing Rain', 'Light Snow and Sleet / Windy', 'Snow and Sleet / Windy', 'Sleet / Windy', 'Heavy Freezing Rain / Windy', 'Squalls / Windy', 'Light Rain Shower / Windy', 'Snow and Thunder / Windy', 'Light Sleet / Windy', 'Sand / Dust Whirlwinds', 'Mist / Windy', 'Drizzle / Windy', 'Duststorm', 'Sand / Dust Whirls Nearby', 'Thunder and Hail', 'Heavy Sleet', 'Freezing Rain / Windy', 'Light Snow Shower / Windy', 'Partial Fog', 'Thunder / Wintry Mix / Windy', 'Patches of Fog / Windy', 'Rain and Sleet', 'Light Snow Grains', 'Partial Fog / Windy', 'Sand / Dust Whirlwinds / Windy', 'Heavy Snow with Thunder', 'Light Snow Showers', 'Heavy Blowing Snow', 'Light Hail', 'Heavy Smoke', 'Heavy Thunderstorms with Small Hail', 'Light Thunderstorm', 'Heavy Freezing Drizzle', 'Light Blowing Snow', 'Thunderstorms and Snow', 'Dust Whirls', 'Rain Shower / Windy', 'Sleet and Thunder', 'Heavy Sleet and Thunder', 'Drifting Snow / Windy', 'Shallow Fog / Windy', 'Thunder and Hail / Windy', 'Heavy Sleet / Windy', 'Sand / Windy', 'Heavy Rain Shower / Windy', 'Blowing Snow Nearby', 'Heavy Rain Shower', 'Drifting Snow'

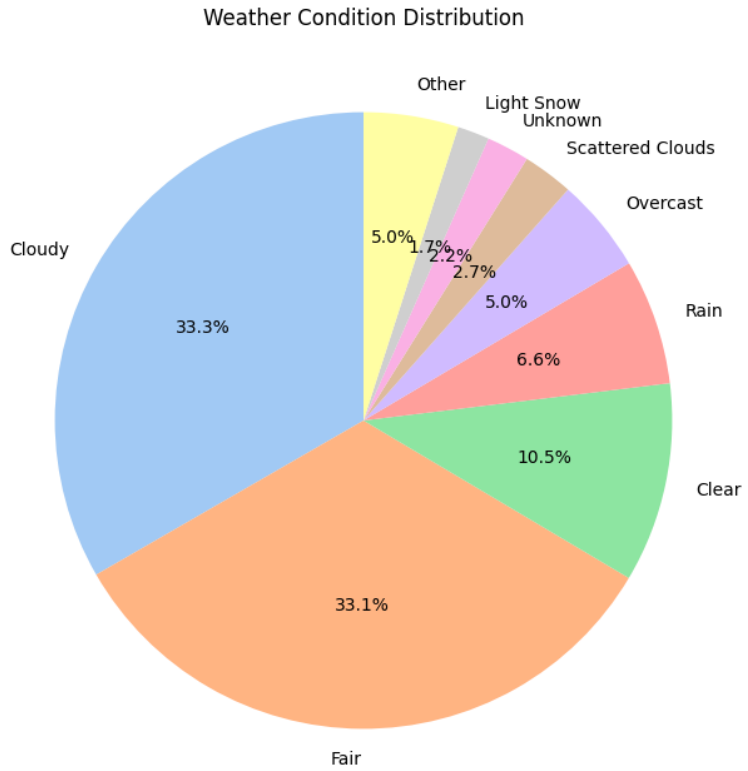

Figure5. Distribution of Weather Condition

표5를 보면 구름낀(Cloudy)상태가 가장 많고 비슷한 비중으로 맑은(Fair)상태가 있다. 뒤이어 다양한 날씨상태의 비중이 나온다. 구름낀 상태에서 사고가 많이 발생하긴 했으나 맑고 화창한 상태에서도 많은 사고가 일어났다. 때문에 어느 특정 날씨 상태에 사고가 더 많이 발생한다고 보기 힘들다.

2.3 도로환경적 측면

사고를 도로환경적 측면에서 접근하기 위해 선택한 변수는 다음과 같다.
'Amenity', 'Bump', 'Crossing', 'Give_Way', 'Junction', 'No_Exit', 'Railway', 'Roundabout', 'Station', 'Stop', 'Traffic_Calming', 'Traffic_Signal', 'Turning_Loop'
해당 변수들은 모두 불리언 값으로 참,거짓 값만 가진다. Countplot을 이용해 해당 변수들의 참,거짓 개수를 세어봤다.

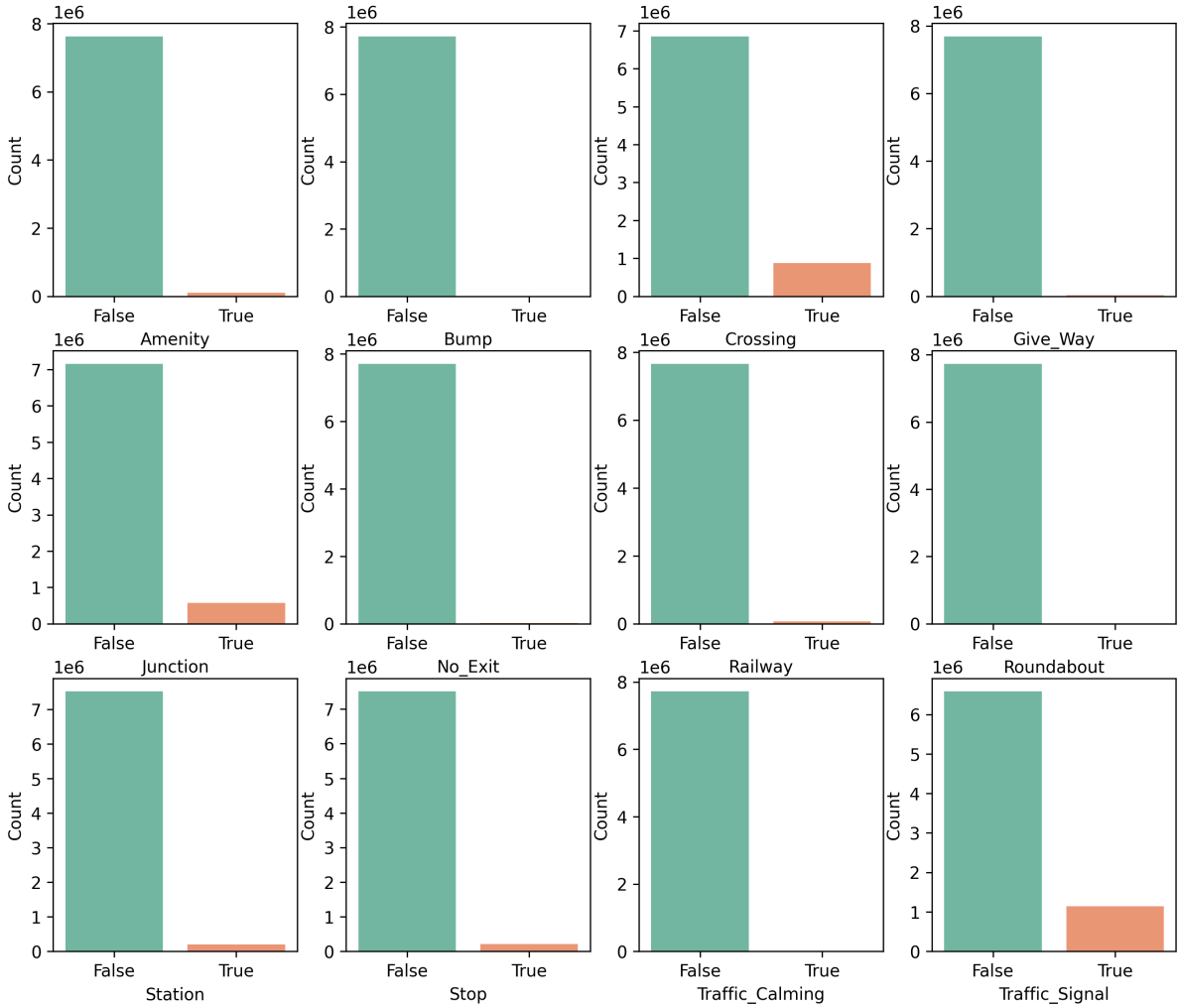

Figure6. Countplot of Road Environment Data

표6을 보면 대부분의 변수에서 거짓이 대부분을 차지하고 있는 것을 볼 수 있다. 횡단보도(Crossing), 교차로(Junction), 신호등(Traffic_Signal) 변수에서만 참 값이 그나마 눈에 띌 정도의 개수를 보이고 있음을 알 수 있다. 이를 통해 횡단보도, 교차로, 신호등이 주변에 있는 상황을 제외하고는 다른 도로환경 요소들이 자동차 사고에 영향을 미치지 않았음을 확인할 수 있다.

3. Multivariate analysis

사고의 심각도에 시간 데이터와 날씨 데이터가 영향을 미치는지 살펴보도록 하겠다.

3.1 사고 심각도와 시간 데이터

시간에 따른 사고 심각도에 변화가 있는지 확인하기 위해 시간 데이터로는 연도, 월, 요일, 시간대를 선택했다. 분석기법으로는 스피어만 상관계수를 사용했다.

	Year	Month	Weekday	Hour	Severity
Year	1.000000	-0.160138	0.084667	0.036122	-0.280904
Month	-0.160138	1.000000	-0.003241	0.020103	-0.012714
Weekday	0.084667	-0.003241	1.000000	0.041974	0.018134
Hour	0.036122	0.020103	0.041974	1.000000	0.014871
Severity	-0.280904	-0.012714	0.018134	0.014871	1.000000

Table1. Spearman Correlation Matrix Between Severity and Time Variables

연도와 월은 음의 상관계수를 가지고 요일와 시간대는 양의 상관계수를 가진다. 이는 연도와 월이 증가함에 따라 사고의 심각도가 감소하고, 요일과 시간대가 증감할 때 사고의 심각도도 증가한다는 사실을 알 수 있다. 하지만 모두 상관계수가 작아 강한 상관관계를 보인다고 보기는 힘들다. 연도가 음의 상관관계를 가지는 것은 2023년 사고 데이터가 다른 연도에 비해 부족하기 때문으로 추측해볼 수 있다.

3.2 사고 심각도와 날씨 데이터

날씨 데이터와 사고 심각도 사이에 상관관계가 있는지 확인하기 위해 날씨 데이터로 Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Speed(mph), Precipitation(in)를 선택했다. 분석기법으로는 스피어만 상관계수를 사용했다.

	Temperature(F)	Wind_Chill(F)	Humidity(%)	Pressure(in)	Visibility(mi)	Wind_Speed(mph)	Precipitation(in)	Severity
Temperature(F)	1.000000	0.999125	-0.345783	0.084817	0.261888	0.098570	-0.134795	-0.017574
Wind_Chill(F)	0.999125	1.000000	-0.342001	0.086825	0.262215	0.078370	-0.138666	-0.019010
Humidity(%)	-0.345783	-0.342001	1.000000	0.085036	-0.470096	-0.204945	0.338087	0.016759
Pressure(in)	0.084817	0.086825	0.085036	1.000000	0.061724	-0.045019	-0.042904	-0.037022
Visibility(mi)	0.261888	0.262215	-0.470096	0.061724	1.000000	0.045062	-0.426306	-0.025156
Wind_Speed(mph)	0.098570	0.078370	-0.204945	-0.045019	0.045062	1.000000	0.072921	0.031336
Precipitation(in)	-0.134795	-0.138666	0.338087	-0.042904	-0.426306	0.072921	1.000000	0.043357
Severity	-0.017574	-0.019010	0.016759	-0.037022	-0.025156	0.031336	0.043357	1.000000

Table.2 Spearman Correlation Matrix Between Severity and Weather Variables

기온(Temperature)과 가시거리(Visibility), 기압(Pressure)은 사고 심각도와 음의 상관관계를 가지며 나머지 변수들은 양의 상관관계를 보인다. 하지만 모든 상관계수가 0.05이하로 매우 약한 상관관계를 가진다고 볼 수 있다. 따라서 날씨에 따른 사고 심각도의 변화는 크지 않을 것으로 해석된다.

전체적으로 봤을 때 사고 심각도에 크게 영향을 미치는 변수는 찾을 수 없었다.

4. Suggestion

자동차 사고 예방방지의 목적으로 볼 때, 표1에서 확인할 수 있듯이 평일 출퇴근시간(7~8시, 16~17시) 그리고 겨울(11월~1월)에 특히 조심해야 함을 알 수 있다. 그리고 표4에서 확인할 수 있듯이 습도가 높을수록 사고가 더 많이 발생한다. 그렇기 때문에 구름이 많이 끼거나, 안개가 있는날, 비 또는 눈이 오는 날 조심해야 할 필요가 있다. 마지막으로 표6에서 확인할 수 있듯이 다른 곳보다는 신호등, 교차로, 횡단보도가 있는 곳에서 사고가 날 확률이 높으니 운전할 때 특히 조심하도록 해야겠다.
만약 경찰에게 제안할 수 있다면 해당 시간대에 감시 인력을 더 많이 배치하고 해당 도로환경에 감시카메라를 설치하거나 안전 표지판을 세우도록 제안할 수 있을 것이다.