Exploratory Data Analysis

- StudentID: 21800379

- Name: 성현수

- 1st Major: 공연영상

- 2nd Major: 데이터사이언스

도로 상황과 기상환경에 따른 사고의 발생 빈도와 심각도를 확인해본다. 일반적으로 알려진 사고 유발의 원인이 되는 기상 환경과 도로의 복잡도보다 운전자의 주행 중 집중도와 주행행태가 사고에 미치는 영향에 집중해보았다.

1. Data overview

1-1. 데이터 설명

2016년 2월 8일부터 2023년 3월 31일까지의 미국 내 교통사고가 기록된 데이터셋이다. 사고 발생 시간, 위치, 심각도 등 사고 정보에 대한 변수들, 당시의 기상 정보에 대한 변수들 그리고 그외 사고 지역의 도시환경에 대한 변수들로 이루어져있다.

- Sample Size : 7,728,394 observations
- Number of variables : 46 variables

1-2. 전처리

- 결측치
    - End_Lat, End_Lng: 절반 이상의 결측치가 존재한다. Distance(mi)가 0인 경우에는 Start_Lat, Start_Lng 값으로 대체하였으나, 그 외의 경우는 결측치를 그대로 두었다.
    - End_Lat: 3,402,762개
    - End_Lng: 3,402,762개
    - Distance(mi)가 0인 경우에는 Start 좌표로 대체했고, 0이 아닌 경우는 결측치를 처리하지 않았다.
    - Precipitation(in): 결측치가 많아 사용하지 않기로 결정했다. 대신 Weather_Condition을 바탕으로 Rainy 변수를 생성해 비 관련 기상 조건과 다른 기상 조건을 구분했다.
- 이상치
    - Temperature(F)와 Wind_Chill(F): 일부 이상치가 확인되었다. 이를 위해 Season(계절) 변수를 추가하여, 각 계절별 IQR(Interquartile Range)을 활용해 이상치를 감지하고 제거했다. 아래는 이상치를 확인하기위한 박스 플롯이다.
- 데이터타입
    - Start_Time과 End_Time: 사고 발생 시간과 종료 시간이 object 형식으로 저장되어 있었으나, 이를 datetime 형식으로 변환했다. 변환 과정에서 초 단위 이하의 소수값들은 모두 제거했다.

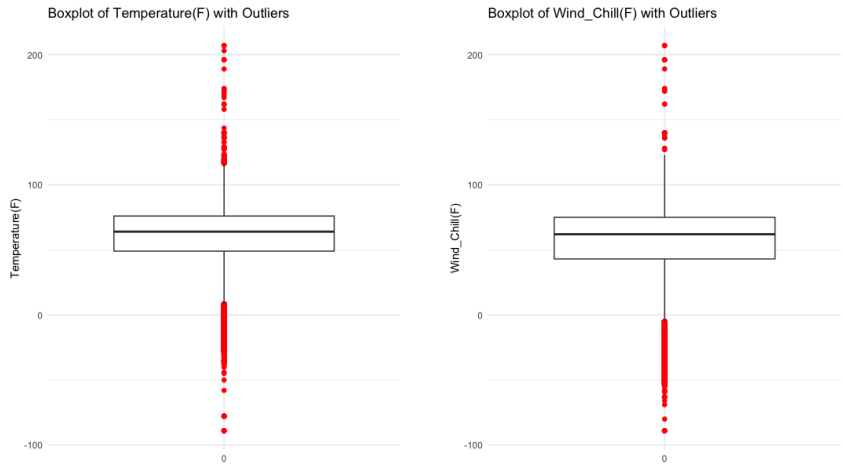

Figure 1. Outliers of Temperature(F) and Wind Chill(F)

1-3. 주요변수

Name	Data type	Description
ID	object	사고 고유 식별자
Severity	int 64	사고의 심각도 (1~4)
Start_Time	datetime64(ns)	사고 발생 시간
End_Time	datetime64(ns)	사고 종료 시간
Start_Lat, Start_Lng	float64	사고 발생 지점의 위경도
End_Lat, End_Lng	float64	사고 종료 지점의 위경도
Distance(mi)	float64	사고로 인해 영향을 받은 도로 구간의 길이
Temperature(F)	float64	사고 당시의 기온
Wind_Chill(F)	float64	사고 당시의 체감 온도
Humidity(%)	float64	사고 당시의 습도
Pressure(in)	float64	사고 당시의 기압
Visibility(mi)	float64	사고 당시의 가시거리
Wind_Speed(mph)	float64	사고 당시의 풍속
Precipitation(in)	float64	사고 당시의 강수량
Weather_Condition	object	사고 당시의 기상 상태
Rainy	object	비 발생 여부 (Rain 혹은 Others)
Is_Weekend	int64	사고 발생 주말/주중 여부 (1:주말, 0:주중)
Time_of_Day	object	사고 발생 시간대 (아침, 낮, 저녁, 밤)
Season	object	사고 발생 계절 (봄, 여름, 가을, 겨울)

Table 1. 주요변수

2. Univariate analysis

2.1 기상 환경에 따른 사고 발생 비율

기상 정보에 따른 사고 발생 비율을 확인하였다. 비와 관련된 기상 정보가 세분화되어 나타나있어 맑은 날씨(Fair, Clear)와 비교하기에 적합하지 않다고 판단하였다. 따라서 Rainy 변수를 새롭게 만들어 Weather_Condition 값 중 'Snow, Rain, Thunder, Squalls, Storm, Hail, Sleet, Showers, Pellet, Drizzle'이 포함된 경우 'rain' 값으로, 포함하지 않은 경우 기존의 값으로 저장하였다. 이후 얻게 된 파이차트는 다음과 같다.

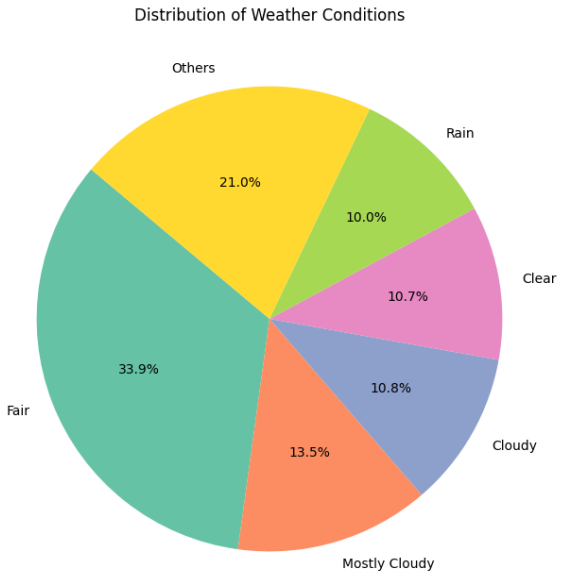

Figure 2. Distribution of Weather Condition

기상 환경이 좋지 않을 경우 사고 발생 위험도가 올라가지만, 실제 집계량을 보았을 때 비 혹은 비와 관련된 기상 현상(눈, 우박, 태풍 등)이 일어날 때에 비해 맑은 날, 흐린 날 등의 사고 건수가 더욱 많았다. 일반적으로 알려진 악천후에서의 사고 위험에 대한 인식으로 인해 오히려 맑은 날씨에 운전자의 주의 집중 및 운전 행태가 사고 위험에 노출되기 쉽게 변할 가능성을 짐작해보았다.

2.2 보행자 교통량

미국 내 보행자 유동인구가 많은 지역과 적은 지역의 사고수를 비교하였다. 보행자의 도로 간섭을 짐작할 수 있는 두 변수 Traffic_Signal, Crossing을 사용하여 두 변수 중 최소 한 개의 True가 있을 경우 보행자 교통량 많음(High Pedestrian Traffic), 모두 False일 경우 보행자 교통량 적음(Low Pedestrian Traffic)을 반환하는 새로운 변수 Pedestrian_Traffic를 생성했다. 보행자가 많은 도로일수록 도로의 복잡도와 차량 통행량도 많을 것으로 예상했다. 결과는 다음과 같다.

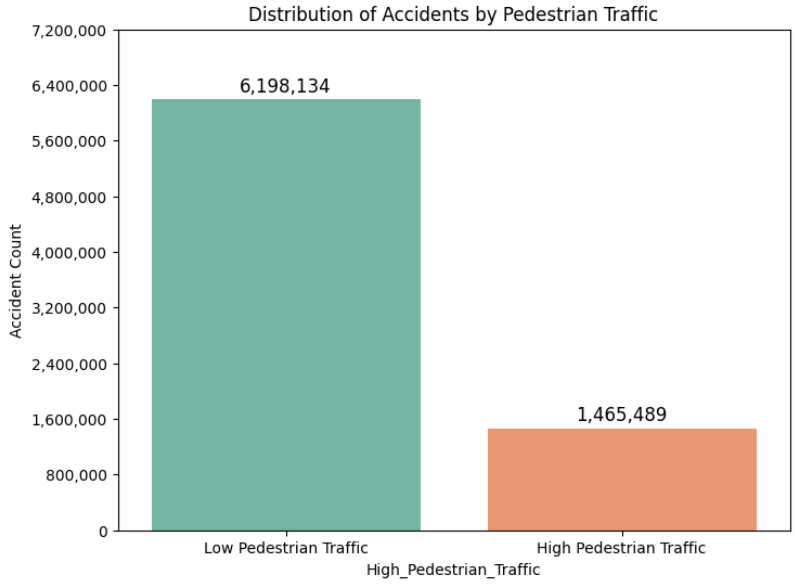

Figure 3. Distribution of Accidents by Pedestrian Traffic

예상과 달리 보행자 교통량이 적은 도로에서의 사고 집계량이 눈에 띄게 큰 것을 확인하였다. 도로의 복잡도 혹은 차량 통행량보다 다른 요인들이 사고유발에 영향을 끼침을 유추해볼 수 있었다.

3. Multivariate analysis

3.1 보행자 교통량에 따른 사고 심각도

보행자 교통량과 도로의 복잡도에 따른 사고 심각도의 분포를 확인하였다. 보행자 교통량와 도로의 복잡도가 극단적으로 차이가 나는 미국의 도로 특성상 사고의 심각도에서도 유의미한 차이를 확인할 수 있을 것으로 생각하였다.

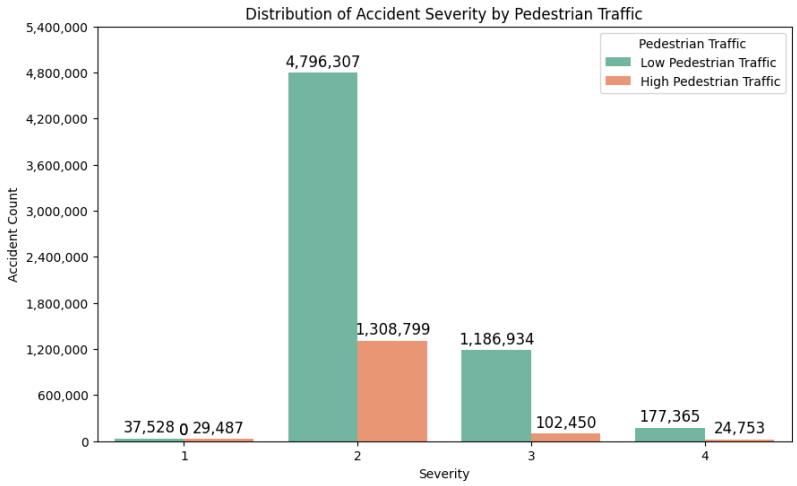

Figure 4. Distribution of Accident Severity by Pedestrian Traffic

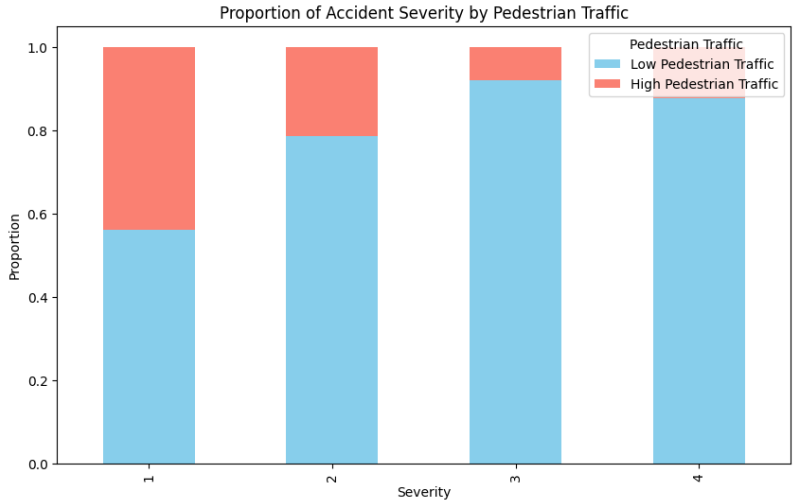

Figure 5. Proportion of Accident Severity by Pedestrian Traffic

보행자 교통량이 많을수록 경미한 사고의 비중이 크고, 사고의 심각도가 높아질수록 보행자 교통량이 적은 (도심지 외곽과 그 밖으로 가정) 지역에서 발생한 사고의 비중이 커지는 경향을 확인할 수 있었다. 앞서 확인한 날씨에 따른 사고 건수와 동일한 맥락으로 도로의 복잡도가 높거나 사고 발생 가능성이 높을수록 운전자가 더욱 주의를 집중하여 운전하기에 사고의 발생 수의 비중과 심각도가 낮은 것으로 유추해보았다.

3.2 사고 심각도별 사고처리 지연시간과 도로막힘 구간길이

사고가 발생한 경우 이로인한 도로 지연이 사회적 손실을 일으킨다. 사고처리 시 소요되는 시간와 도로 막힘의 영향을 받은 구간의 길이를 도심지(보행자 교통량이 많은 경우)와 외곽 및 그 외 지역(보행자 교통량이 적은 경우)로 나누어 사고 심각도별로 확인해보았다.

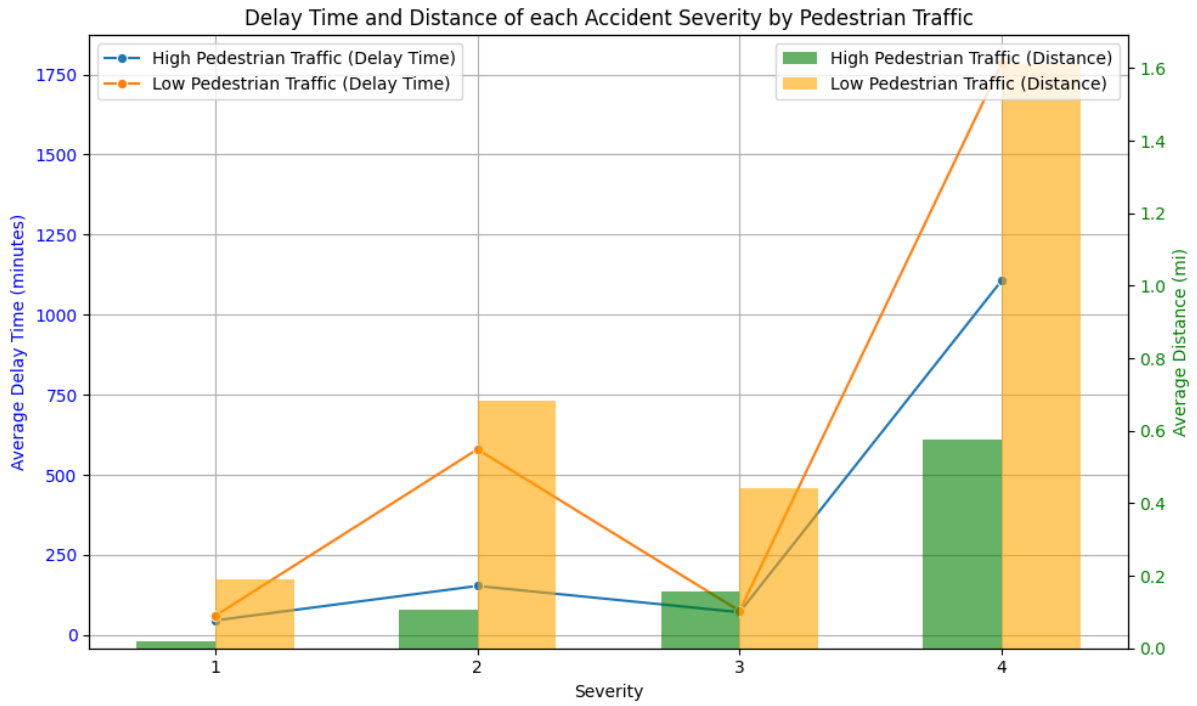

Figure 6. Delay Time and Distance of each Accident Severity by Pedestrian Traffic

도로 외곽 및 그 외 지역의 사고처리 지연시간과 도로막힘 구간길이가 도심지에 비해 더욱 긴 것으로 확인된다. 우회도로도 제한적일 뿐더러 초동 조치를 위한 접근성 또한 떨어지는 것이 그 원인인 것으로 판단된다.

4. Suggestion

비교적 사고 위험도가 낮은 것으로 판단되는 기상환경과 도로 상황에 따른 운전자의 부주의로 인해 발생하는 사고가 큰 것으로 확인된다. 이로 인한 사고의 조치 또한 어려운 것을 확인하였다. 도심지 외곽과 그 외 지역의 사고 초동 조치를 위한 시스템을 재정비할 필요성과 함께 주행 중 주의 집중을 위한 장치들에 대한 고민이 필요해보인다. 사고가 빈번히 발생하는 지역의 좌표 정보를 통해 신속한 사고 초동 조치를 위한 시스템 구축과 운전자를 주의시키기 위한 도로 재정비 등을 위한 추가적인 연구가 기대된다.