Exploratory Data Analysis

- StudentID: 22000415

- Name: 양찬

- 1st Major: ACE

- 2nd Major: DS

다음과 같은 EDA 체크리스트를 반영해 EDA를 진행하고자 한다.

[EDA 체크리스트]

1. *어떤 질문을 풀거나 틀렸다고 증명하려고 하는가?*
2. *중복된 항목이 있는가?*
3. *어떤 종류의 데이터가 있으며 데이터 타입들을 어떻게 다루려고 하는가?*
4. *데이터에서 누락된 것이 있는지, 있다면 그것들을 어떻게 처리하려는가?*
5. *이상치는 어디에 있는가? 관심을 가져야 할 데이터인가?*
6. *변수 간 상관성이 있는가?*
    1.

이번 연구의 목적은 교통사고 데이터 분석을 통해 사고의 심각도에 영향을 미치는 주요 환경 요인(기상, 도로 조건 등)과 시간대, 지역별 사고 발생 패턴을 파악하는 것이다. 이를 통해 사고 예방을 위한 효과적인 교통 관리 및 정책 수립에 기여하고자 한다.

1. Data overview

2. 중복된 항목이 있는가? : 없음
3. 어떤 종류의 데이터가 있으며 데이터 타입들을 어떻게 다루려고 하는가?

- Data size : 7728394 X 46

Name	Type	Range	Unique	Description
ID	object	A-1 ~ A-7777761	7728394	Unique identifier for each accident record.

Name	Type	Range	Unique	Description
Source	object	Source1, Source2, Source3	3	The source from which the accident information is obtained.
Severity	int64	1, 2, 3, 4	4	A numerical value representing the severity of the accident
Start_Time	object	2016-02-08 05:46:00 ~ 2019-08-23 18:52:06	6131796	The date and time when the accident occurred.
End_Time	object	2016-02-08 11:00:00 ~ 2019-08-23 19:21:31	6705355	The date and time when the accident was resolved or no longer affecting traffic.
Start_Lat	float64	39.063148, ... ,34.120911	2428358	The latitude where the accident started.
Start_Lng	float64	-84.032608, ... ,-118.416176	2482533	The longitude where the accident started.
End_Lat	float64	39.86501, ... ,33.943599	1568172	The latitude where the accident ended
End_Lng	float64	-84.04873, ... ,-118.416176	1605789	The longitude where the accident ended
Distance(mi)	float64	1.3200e+00, ... ,1.1622e+01	22382	The distance in miles affected by the accident.
Description	object	-	3761578	A textual description of the accident.
Street	object	State Route 32, ... ,N Digital Dr	336306	The street where the accident occurred.
City	object	Williamsburg, ... ,Ness City	13678	The city where the accident occurred.
County	object	Clermont, ... ,Woods	1871	The county where the accident occurred.
State	object	OH, ... ,CH	49	The state where the accident occurred.
Zipcode	object	45176, ... ,97827-9687	825094	The postal code of the accident location.
Country	object	'US'	1	The country of the accident.

Name	Type	Range	Unique	Description
Timezone	object	US/Eastern, US/Pacific, nan ,US/Central, US/Mountain	4	The time zone of the accident location.
Airport_Code	object	KI69, ... ,KY22	2045	The nearest airport code to the accident location.
Weather_Timestamp	object	2016-02-08 05:58:00 ~ 2019-08-23 01:20:00	941331	The timestamp when the weather information was recorded.
Temperature(F)	float64	3.690e+01, 3.790e+01, ...	860	The temperature at the time of the accident in Fahrenheit.
Wind_Chill(F)	float64	31. , ... ,-65.9	1001	The wind chill factor at the time of the accident in Fahrenheit.
Humidity(%)	float64	1 ~ 100	100	The humidity percentage at the time of the accident.
Pressure(in)	float64	29.67, ... ,31.13	1144	The air pressure at the time of the accident in inches.
Visibility(mi)	float64	1.00e+01, 9.00e+00, ...	92	The visibility in miles at the time of the accident.
Wind_Direction	object	Calm, ... , SSW	24	The direction of the wind at the time of the accident.
Wind_Speed(mph)	float64	0 ~ 10	184	The wind speed in miles per hour at the time of the accident
Precipitation(in)	float64	0 ~ 10	299	The amount of precipitation in inches at the time of the accident.
Weather_Condition	object	Scattered Clouds, Partly Cloudy , ...	144	The general weather conditions at the time of the accident
Amenity	bool	T/F	2	Indicates if the accident occurred near an amenity
Bump	bool	T/F	2	Indicates if the accident occurred near a bump
Crossing	bool	T/F	2	Indicates if the accident occurred near a crossing

Name	Type	Range	Unique	Description
Give_Way	bool	T/F	2	Indicates if the accident occurred near a give-way sign
Junction	bool	T/F	2	Indicates if the accident occurred near a junction
No_Exit	bool	T/F	2	Indicates if the accident occurred near a no-exit sign
Railway	bool	T/F	2	Indicates if the accident occurred near a railway
Roundabout	bool	T/F	2	Indicates if the accident occurred near a roundabout
Station	bool	T/F	2	Indicates if the accident occurred near a station
Stop	bool	T/F	2	Indicates if the accident occurred near a stop sign
Traffic_Calming	bool	T/F	2	Indicates if the accident occurred near a traffic calming measure
Traffic_Signal	bool	T/F	2	Indicates if the accident occurred near a traffic signal
Turning_Loop	bool	T/F	2	Indicates if the accident occurred near a turning loop
Sunrise_Sunset	object	Night, Day, nan	2	The time of day at the time of the accident
Civil_Twilight	object	Night, Day, nan	2	The civil twilight phase at the time of the accident
Nautical_Twilight	object	Night, Day, nan	2	The nautical twilight phase at the time of the accident
Astronomical_Twilight	object	Night, Day, nan	2	The astronomical twilight phase at the time of the accident

- 시간(Start_Time)과 장소(State)에 따른 데이터 추세 파악
    - Start_Time을 기준으로 시각화
        - Start_Time, End_Time 중 결측치가 적은 Start_Time 선택

    - 장소(State)에 따른 빈도를 살펴보고 주목할만한 특징이 있는지 파악
        - 장소 관련 변수가 많은데 유니크값이 가장 적은 state를 선택

- 범주화할 수 있는 변수들을 먼저 전처리
    - Severity, State, Timezone : 유니크 값이 적어서 범주형으로 변환 가능
    - Precipitation(in) : 미국 강우량 강도 분류 지표에 따른 범주화 가능 (https://www.baranidesign.com/faq-articles/2020/1/19/rain-rate-intensity-classification)
- Key_variable : Severity(사고 심각도), 사고 빈도 와의 관계성
    - 이전 연구를 바탕으로 이들에게 가장 영향이 있을 것 같은 변수들과의 관계부터 시각화한다.
    - Weather_Condition, Visibility(mi), Precipitation(in), Wind_Speed(mph), Traffic_Signal, Temperature(F)

4. 데이터에 누락된 값은 없는지, 있다면 그것을 어떻게 처리하려는가? : 없다

5. 이상치는 어디에 있는가? 관심을 가져야 할 데이터인가?

- 이상치가 있는 칼럼 :

Name	Number of NA
End_Lat	3402762
End_Lng	3402762
Street	10869
City	253
Zipcode	1915
Timezone	7808
Airport_Code	22635
Weather_Timestamp	120228
Temperature(F)	163853
Wind_Chill(F)	1999019
Humidity(%)	174144
Pressure(in)	140679
Visibility(mi)	177098
Wind_Direction	175206
Wind_Speed(mph)	571233
Precipitation(in)	2203586
Weather_Condition	173459

Name	Number of NA
Sunrise_Sunset	23246
Civil_Twilight	23246
Nautical_Twilight	23246
Astronomical_Twilight	23246

- 주목하기로 한 변수들 중 Visibility(mi)는 급박한 사고 상황에서 누락되기 쉽기 때문에 생긴 NA 값이라 생각하며 Precipitation(in), Wind_Speed(mph), Traffic_Signal, Temperature(F) 는 수치형 변수들로 각각의 항목에 데이터가 해당되지 않을수 있기 때문에(ex. 비가 안오는 날은 강수량이 없음.) 생긴 NA 값이라고 볼 수 있을 것이다.

2. Univariate analysis

Presentation of key variables from various aspects

2.1 Location

먼저 장소 변수(Start_Lat, Start_Lng, States)를 기준으로 시각화를 진행해보았다.

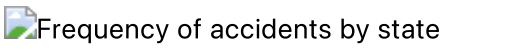
Figure 1. Frequency of accidents by state

이 그래프는 주별 사고 발생 빈도를 보여준다. 캘리포니아(CA), 플로리다(FL), 텍사스(TX)가 사고 발생이 가장 많다. 이 주들은 인구가 많고 교통량도 많아서 그럴 것이라고 예상된다. 이에 반해 사우스 다코타(SD)와 버몬트(VT)는 인구가 가장 적은 주 중 일부이며 그렇기에 사고가 가장 적게 발생하고 있는 것을 확인할 수 있다.


Figure 2. Top 5 states by Severity

이 그래프는 Start_Lat, Start_Lng 를 기준으로 사고 발생 위치를 찍은 산점도로 Severity(사고심각도)가 클수록 더 진한색으로 표현되도록 했다.

End가 아닌 Start_Lat, Start_Lng 를 기준으로 그린 이유는 End_Lat, End_Lng 은 3402762개의 결측치가 존재하기 때문에 더 정확한 결과를 위해 Start_Lat, Start_Lng 을 기준으로 그래프를 그렸다.

또한 사고 빈도가 높은 상위 5개 주의 위치를 파란원으로 보여주는데 CA(캘리포니아), MI(미시간), GA(조지아), SC(사우스캐롤라이나)가 그 예시이다. 캘리포니아(CA)에 두개의 주요 스팟이 형성된 것으로 보아 CA가 서부에서 가장 사고가 빈번한 곳이란 것을 알 수 있으며 나머지 스팟 모두 동남부 지역인 것으로 보아 북부를 제외한 서부와 동남부 지역의 사고 예방 조치가 특히 필요할 것으로 보인다.

2.2 Time

두번째는 시간 변수(Start_Time)를 기준으로 시각화해보았다.

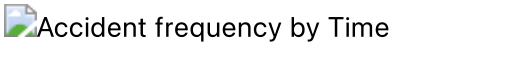
Figure 3. Accident frequency by Time

Start_Time 의 값은 2016-02-08 05:58:00 형식으로 이루어져 시간 정보를 모두 가지고 있는 변수이다. 이것을 Year, Month, Hour, Weekday 로 나누어 분포를 확인해보았다.

1. Accidents by Year :

- 2016년부터 2023년까지의 범주를 가지고 있으며 16 ~ 19 까지는 증가폭이 완만하다가 20년도부터 증가폭이 커지기 시작하더니 21년에 최대 폭으로 사고 발생이 늘어난다.
- 이는 2020년 초 COVID-19 팬데믹으로 인해 전 세계적으로 이동이 제한되어 교통량이 줄었다가 2020년 하반기부터 백신 개발과 규제 완화가 이루어지면서 교통량이 급격히 증가했고 갑작스러운 교통량 증가는 사고 발생률을 높이는 주요 원인이 되었을 것이다. 이러한 경향은 22년까지 지속되어 사고가 증가하였다.
- 2023년의 경우 아직 연말까지 데이터가 수집되지 않아 가장 적은 양을 보이는 것으로 보인다.

2. Accidents by Month :

- 월별 교통사고 발생 빈도를 보여주는 그래프로, 12월에 가장 많은 사고가 발생했으며, 6월에 사고가 가장 적게 발생했다.
- 사고 발생 빈도가 연말(11월, 12월)에 증가하는 경향을 보이는데, 이는 겨울철 기상 조건(눈, 빙판길 등)이 악화되면서 사고 위험이 증가했기 때문일 수 있다.

3. Accidents by Hour :

- 하루 중 시간대별 사고 발생 빈도를 나타낸 그래프로, 아침과 저녁 출퇴근 시간대(7시~9시, 16시~18시)에 사고가 집중되는 경향을 보인다.
- 특히, 오전 7시9시와 오후 16시18시에 사고가 가장 빈번하게 발생하는 것을 확인할 수 있다. 이는 출퇴근 시간대 교통량이 집중되는 현상과 밀접한 관련이 있다.
- 새벽 시간대(0시~5시)에는 사고 빈도가 현저히 적고, 하루가 끝나갈수록(22시~24시) 다시 사고 발생이 감소하는 경향을 보인다.

4. Accidents by Weekday :

- 요일별 사고 발생 빈도를 나타낸 그래프로, 월요일부터 금요일까지는 사고 빈도가 비슷하게 유지되며, 주중 내내 교통량이 높다가 주말(토요일과 일요일)에는 사고 빈도가 크게 줄어든다. 이는 주말에는 출퇴근이 없기 때문에 교통량이 줄어들 가능성이 크다.

2.3 Variables that are likely to affect (Visibility, Precipitation)


Figure 4. Percentage by category of Visibility & Precision

3. Multivariate analysis

Presenation of hidden patterns between variables (correlation, clustering, etc.)

3.1 Anova Analysis of Accident Severity by Each Time Category

운전자에게 교통 정보를 제공하는 Drivemode 는 Rush-Hour 운전 트렌드를 자세히 살펴보고 출퇴근의 부담이 가장 큰 시기와 장소를 파악하기 위해 시간대를 다음과 같이 분류했다.

(https://www.drivemode.com/blog/engineering/drivemode-data-report-commuting-durations)

- 출근 시간(Morning Rush) (6:00 AM ~ 9:00 AM): 주요 도시들에서는 출근 시간이 이 시간대로 정의되며, 대부분의 사람들이 이 시간에 출근길에 나섭니다. 뉴욕이나 로스앤젤레스 등의 대도시에서 7:00 AM ~ 9:00 AM이 출근 교통이 가장 혼잡한 시간으로 보고되고 있다.

- 낮 시간(Daytime) (9:00 AM ~ 4:00 PM): 이 시간대는 대부분의 직장인들이 근무하는 시간으로, 교통량이 출퇴근 시간에 비해 상대적으로 적다.

- 퇴근 시간(Evening Rush) (4:00 PM ~ 7:00 PM): 퇴근 시간은 보통 4:00 PM부터 시작해 7:00 PM까지 이어지며, 이 시간 동안 교통 혼잡이 심해진다.

- 야간 시간(Night) (7:00 PM ~ 6:00 AM): 퇴근 시간 이후부터 다음날 아침까지는 야간 시간으로 간주되며, 이 시간대에는 상대적으로 교통량이 적다.

ANOVA는 종속변수가 범주형일때, 여러 집단의 평균 차이를 동시에 비교할 수 있다는 장점이 있으므로 이 통계분석 기법을 선택하였다. ANOVA를 사용해 각 시간대별 집단의 평균을 비교하며 종속변수인 Severity에 미치는 영향이 유의한지 파악할 수 있다.

Correlation Matrix
Table 1. Anova Analysis Results

- 결과 : F-값이 1293.255476로 매우 크고 p-값이 0.0으로 0.05보다 작기 때문에, 시간대에 따라 사고 심각도가 유의미하게 차이가 난다고 해석할 수 있다.
- 즉, 출근 시간, 낮 시간, 퇴근 시간, 야간 시간 중 하나의 시간대에서 사고의 심각도가 다를 가능성이 매우 높은데 이는 사후분석을 통해 추가적으로 분석할 수 있다.

3.1-1 Tukey's HSD

Tukey's HSD
Table 2. Tukey's HSD Results

- 종합 결론 : 모든 시간대 쌍 사이에서 p-값이 0.05보다 작기 때문에 사고 심각도 차이가 유의미하게 나타났다. 즉, 출근 시간, 퇴근 시간, 낮 시간, 야간 시간 모두 사고 심각도가 서로 다르며, 특정 시간대에는 더 심각한 사고가 발생할 가능성이 높다.
- 이러한 결과는 교통 혼잡도나 도로 조건, 야간 운전의 위험성 등을 고려한 추가적인 분석과 대책이 필요할 수 있음을 시사한다.

3.2 Correlation

6. 변수 간 상관성이 있는가?

Correlation Matrix
Figure 5. Correlation Matrix

상관관계 그래프를 보면 종속변수인 Severity 와 다른 변수들과의 상관계수는 모두 0.3보다 낮다. 그러나 Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Precipitation(in) 과 같은 변수들은 일반적으로 사고 심각도에 영향을 미칠 수 있는 변수들이기 때문에 이들을 회귀 모델에 포함하는 것이 옳다고 생각했다.

종합했을때 상관계수가 낮은 비선형적 관계를 보여주는 변수들이 있으며, 종속변수가 1,2,3,4 의 4가지 범주를 가진 범주형 데이터이기 때문에 다중 로지스틱 회귀 분석을 하는 것이 적절하다고 판단했다.

3.2-1 Multinomial Logistic Regression


Multinomial Logistic Regression
Table3. Result Matrix of Multinomial Logistic Regression

- 가시거리(Visibility): 가시거리가 낮을수록 심각한 사고(Severity 4)로 분류될 가능성은 낮아지는 경향을 보이며, 전반적으로 사고 심각도에 미치는 영향은 크지 않다.
- 습도(Humidity): 습도는 사고 심각도에 미치는 영향이 매우 작고, 유의미한 변수로 작용하지 않는 경향을 보인다.
- 기압(Pressure): 기압이 높을수록 낮은 심각도의 사고로 분류될 가능성이 높아지며, 특히 Severity 1에서 기압은 중요한 변수로 작용한다.
- 온도(Temperature): 온도가 높을수록 낮은 심각도의 사고로 분류될 가능성이 높으며, 특히 Severity 1에서는 온도가 중요한 변수로 나타난다.
- 강수량(Precipitation): 강수량이 많아질수록 심각한 사고로 분류될 가능성이 커지며, 특히 Severity 3에서 강하게 영향을 미치는 변수이다.

Precipitation (강수량)과 Pressure (기압)는 다른 변수들에 비해 사고 심각도에 상대적으로 큰 영향을 미친다. 특히 강수량은 심각도 Severity 3에서 중요한 변수로 작용한다. 반면, Visibility (가시거리)와 Humidity (습도)는 사고 심각도에 미치는 영향이 상대적으로 작으며, 변수들의 영향은 전반적으로 작거나 중간 정도이다.

전체적으로 계수가 작게 나오는 이유는 결측치를 모두 제거했기 때문으로 보이며 결측치를 변수별 분포에 맞게 대체하는 방법을 사용한다면 더 좋은 결과값이 나올 것이라고 생각한다.

4. Suggestion

- 교통사고 예방을 위한 날씨 정보 활용:

  강수량(Precipitation)과 기압(Pressure)은 사고의 심각도에 큰 영향을 미친다. 특히 강수량이 많을수록 사고가 더 심각해지는 경향이 있으므로, 기상청의 실시간 강수량 데이터를 활용하여 운전자에게 주의 경보를 발령하거나, 비가 많이 오는 날에는 교통 규제를 강화하는 정책을 고려할 필요가 있다.

- 시간대별 교통 관리:

  시간대별 사고 분석 결과, 출퇴근 시간대(Morning Rush, Evening Rush)에 사고 발생 빈도와 심각도가 높다. 출퇴근 시간대에 맞춰 교통 신호나 속도 제한을 조정하거나, 안전요원을 집중 배치하는 정책을 시행하면 사고 감소에 기여할 수 있을 것이다.

- 위험 지역에 대한 집중 관리:

  주별 사고 발생 빈도를 통해 캘리포니아(CA), 미시간(MI), 조지아(GA) 등 사고 발생 빈도가 높은 지역을 파악할 수 있었다. 이 지역들은 교통사고 예방을 위한 안전 조치를 강화할 필요가 있으며, 특히 북부를 제외한 미국 동서부 지역에 대해 집중적인 안전 점검이나 도로 환경 개선이 필요할 것이다.

- 가시거리 및 기압의 영향:

가시거리(Visibility)가 낮을수록 심각한 사고 발생 가능성은 감소하는 경향이 있지만, 여전히 가시거리가 낮은 환경에서 사고가 발생할 가능성이 있기 때문에 안개나 시야가 제한되는 날씨 조건에 맞춘 추가적인 주의 경보 시스템을 마련해야 한다. 또한, 기압이 낮을 때 심각한 사고 발생 위험이 높아지므로, 기상 정보를 활용한 교통 정책을 마련할 수 있다.

- 데이터 품질 및 결측치 처리:

  결측치를 제거하고 분석을 진행하면서 변수가 사고에 미치는 영향이 예상보다 작게 나타났을 수 있다. 향후 분석에서는 결측치를 보다 효과적으로 대체하거나, 결측치 자체가 특정 변수에 영향을 미치는 패턴을 분석해 개선할 수 있을 것이다.