Exploratory Data Analysis

- StudentID: 22100446

- Name: 염지윤

- 1st Major: 경영학

- 2nd Major: 데이터사이언스

Brief summary of your proposed project idea.

- 데이터프레임 컬럼에 대한 이해
- 5개로 나눈 컬럼 집단에서 각각 1개의 Key Variable을 선정하여 Univariate Analysis 수행
- 날씨 변수와 사고 심각성과의 관계 파악을 위한 Multivariate Analysis 수행
- 사고지점 인근 환경 변수와 사고 심각성과의 관계 파악을 위한 Multivariate Analysis 수행
- 분석 결과를 이용한 Suggestion

1. Data overview

Descriptives statistics on overall data

File: accidents_data.csv

	ID	Severity	Description	Start_Time	Distance(mi)	Street	Temperature(F)	Precipitation(in)	Weather_Condition	Bump	Crossing	Railway	Sunrise_Sunset
2208731	A-2218597	3	Left lane blocked due to accident on I-77 Northbound between I-485 and Nations Ford Rd.	2019-03-14 08:43:55	1.49	I-77 N	53.1	NaN	Scattered Clouds	False	False	False	Day

Image 1. Sample of the Accidents Dataframe

Data Set

- 7728394개 행
- 46개의 컬럼
- 행: 교통사고 수
- 컬럼: 자체적으로 5개 그룹으로 구분

컬럼 구분

1. **교통 사고 설명 및 심각성**

- ID, Source, Severity, Description

2. **교통 사고의 시간과 공간**

- Start_Time, End_Time, Start_Lat, Start_Lng, End_Lat, End_Lng, Distance(mi), Street, City, County, State, Zipcode, Country, Timezone, Airport_Code

3. **교통 사고 당시 기후**

- Weather_Timestamp, Temperature(F), Wind_Chill(F), Humidity(%), Pressure(in), Visibility(mi), Wind_Direction, Wind_Speed(mph), Precipitation(in), Weather_Condition

4. **교통 사고 지점 인근 환경**

- Amenity, Bump, Crossing, Give_Way, Junction, No_Exit, Railway, Roundabout, Station, Stop, Traffic_Calming, Traffic_Signal, Turning_Loop

5. **사고 당시 낮과 밤 구분**

- Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight

주요 변수의 데이터 타입

```
ID                   object
Severity              int64
Description          object
Start_Time           object
Distance(mi)        float64
Street               object
Temperature(F)      float64
Precipitation(in)   float64
Weather_Condition    object
Bump                   bool
Crossing               bool
Railway                bool
Sunrise_Sunset       object
```

Image 2. Data Type of Key Variables

2. Univariate analysis

Presentation of key variables from various aspects

2.1 Severity (심각도)

- 1,2,3,4로 구성
- 값 증가: 교통사고 심각도 증가
- Severity 2: 가장 많은 비중 (79.7%)
- Severity 3: 두번째로 많은 비중 (16.8%)

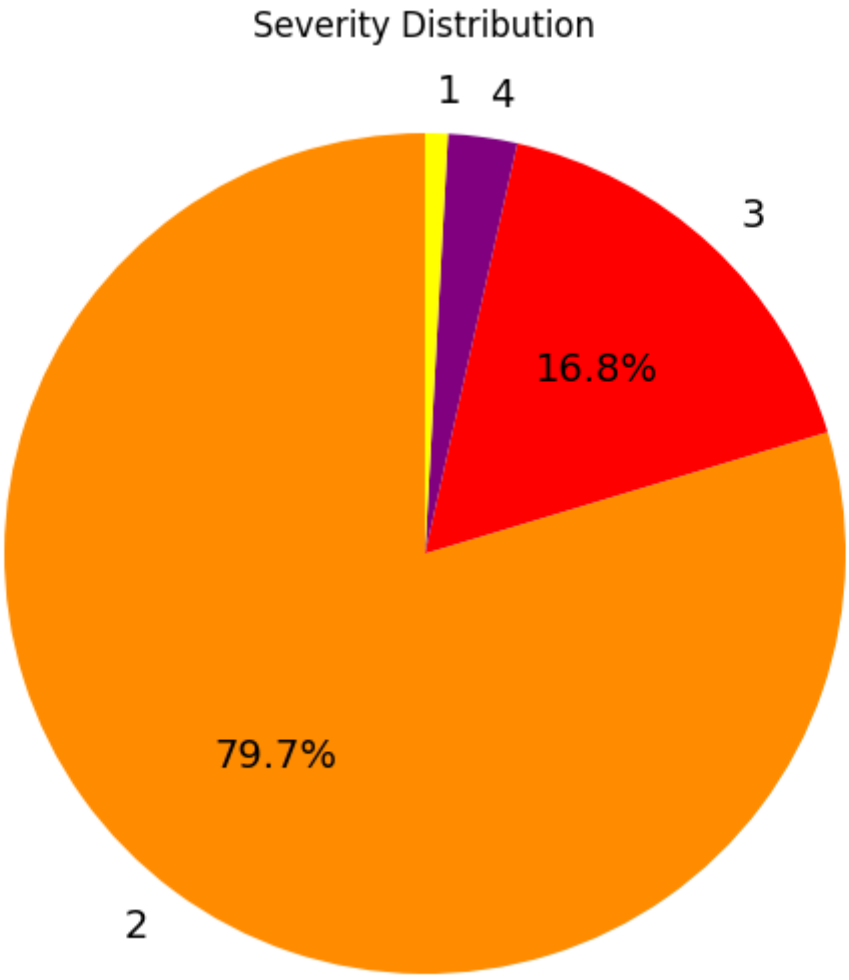

Figure 1. Distribution of Severity

2.2 Bump (과속방지턱)

- True, False로 구성
- Bump 존재: True
- False (99.95%) > True (0.05%)

2.3 Precipitation(in) (강수량)

- 강수량 구간: ['0-0.1', '0.1-1', '1-5', '5-10', '10-20', '20+']
- 구간별 교통사고 발생 빈도
- 0-0.1 구간에서 교통사고 최다 발생

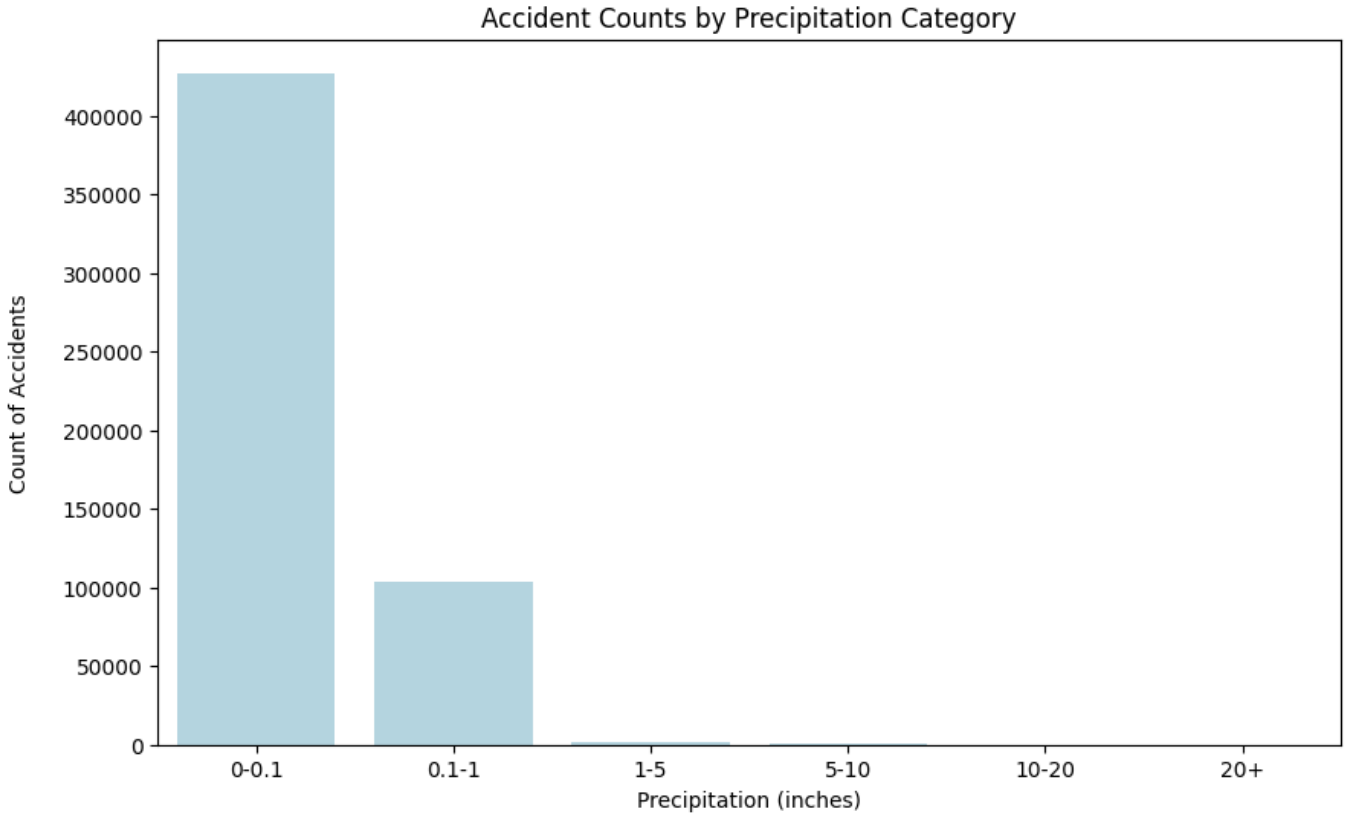

Figure 2. Accident Counts by Precipitation Category

2.4 Sunrise_Sunset (일출 및 일몰 상태)

- Day, Night로 구성
- Day (69.2%) > Night (30.8%)

2.5 Street (사고 발생 거리 정보)

- 교통사고 발생 거리 TOP3
  - I-95 N (78,430회)
  - I-95 S (74,528회)
  - I-5 N (71,968회)

3. Multivariate analysis

Presenation of hidden patterns between variables

3.1 날씨 변수의 사고 심각성 영향 분석

- p-value를 이용한 유의미성 도출
- p-value: 통계적 가설 검정 지표, 0.05 이하 유의미
- 7개 날씨 변수와 심각성(Severity) 관계 분석
- statsmodels 라이브러리로 선형 회귀 모델 적용
- p-value를 -log10(p-value)로 변환
- 큰 값: 심각성에 미치는 영향 통계적 유의미

- 빨간 점선 위 변수: p-value < 0.05

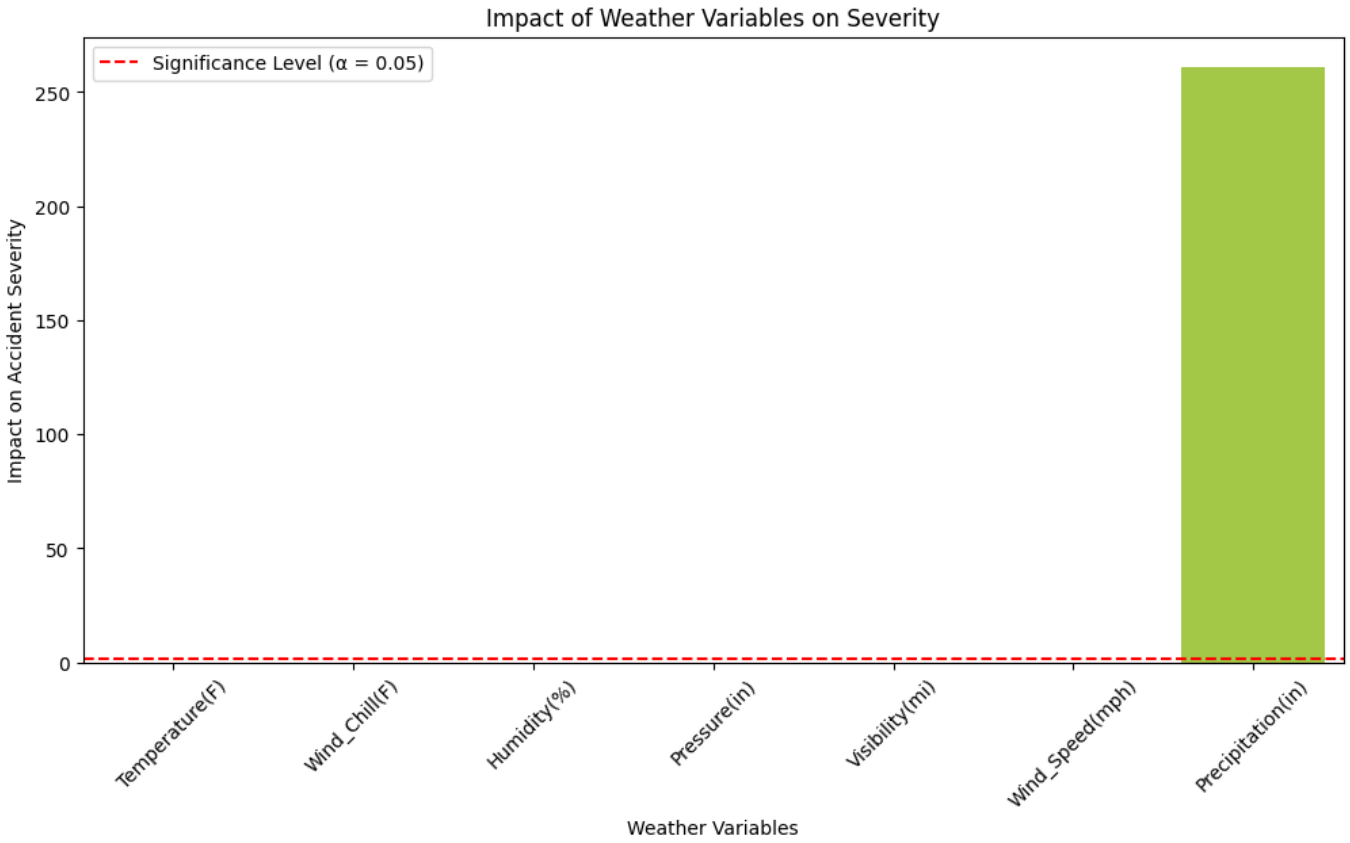

Figure 3. Impact of Weather Variables on Severity

결과 분석

- Precipitation(강수량)만이 사고 심각성과 유의미한 관계 존재
- 다른 날씨 변수: 사고 심각성과 유의미한 관계 없음

3.2 강수량 구간에 따른 사고 심각성 분석

- 3.1 결과에 따른 강수량과 사고 심각성의 유의미한 관계 분석
- 구간별 강수량과 사고 심각성을 박스 플롯으로 분석

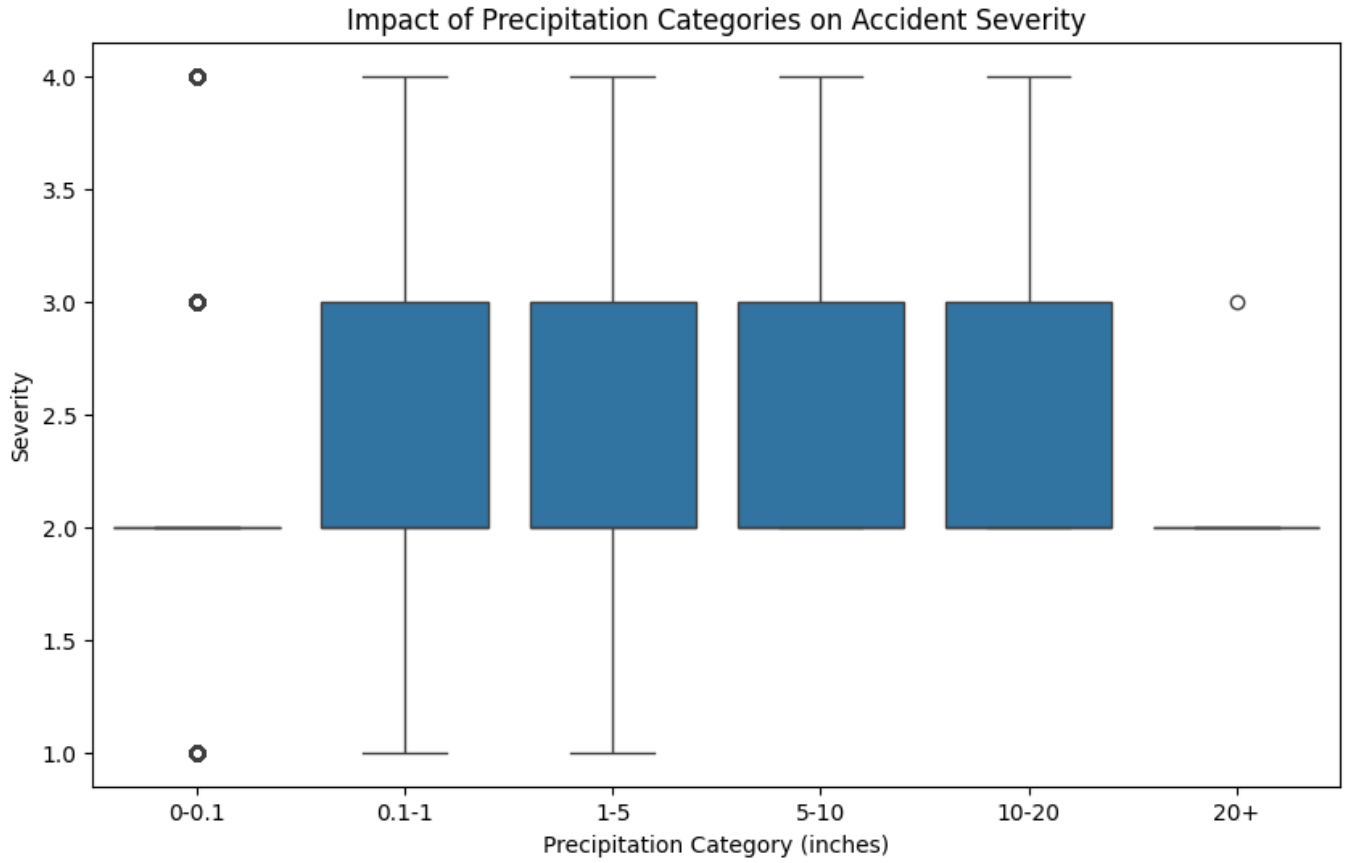

Figure 4. Impact of Precipitation on Accident Severity

결과 분석

- 대부분의 교통사고가 발생한 강수량 0-0.1 구간: 박스 거의 없음
- 0-0.1 구간에서 사고 발생 시 심각성 낮음

3.3 사고 지점 인근 환경 변수의 사고 심각성 영향 분석

- 피셔의 정확 검정 수행
- 13개 인근 교통 환경 변수와 심각성(Severity) 관계 분석
- 각 환경 변수에 대해 Severity와 2x2 교차표 생성
- 피셔 함수를 사용하여 p-value 계산
- p-value를 -log10(p-value)로 변환
- 큰 값: 심각성에 미치는 영향 통계적 유의미
- 빨간 점선 위 변수: p-value < 0.05

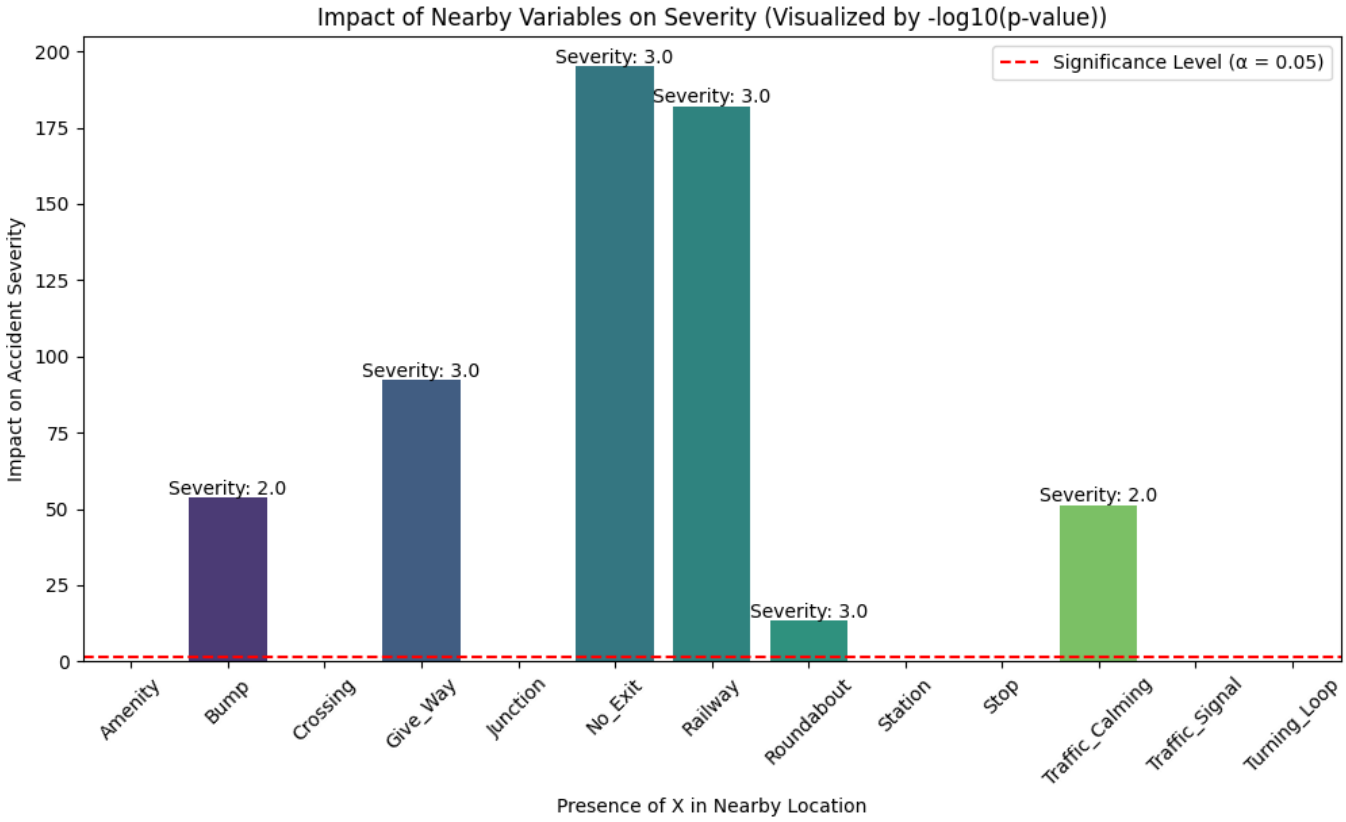

Figure 5. Impact of Nearby Variables on Accident Severity

결과 분석

- No_Exit(도로 끝, 출구 없음)과 Railway(철도)의 존재 여부: 사고 심각성과 유의미한 관계
- Severity가 3일 때, No_Exit와 Railway: 가장 작은 p-value

Severity = 3
-log10(p-value) = 195.04

	Not Severe	Severe
No Exit	6411347	1297502
Exit	17710	1835

Table 1. Contingency Table for No_Exit and Severity

- Severity가 3일 때 No_Exit와 Severity 간의 교차표
- 각 셀의 값: 해당 조건을 만족하는 사고 수 의미

결과 해석

- No Exit: 사고 수 많고 Not Severe 비율 높음
- No Exit 상황: 사고 발생 빈도 높지만, 사고가 상대적으로 심각하지 않음
- Exit: 사고 수 적지만, 심각한 사고 비율 상대적으로 높음
- Exit 상황: 사고 발생 시 보다 심각한 결과 초래 가능성

4. Suggestion

Based on the insights you obtained from the previous stages, propose the potential project idea.

Data Analysis Review

- 분석 큰 틀: 날씨와 도로 교통 장치가 교통사고에 미치는 영향에 초점
- 날씨 변수 중 강수량: 교통사고 심각성과 유의미한 관계
- 사고 발생 관련 강수량: 0-0.1(in), 강수량이 심각성에 미치는 영향 미미
- 교통 환경 변수 중 No_Exit(도로 끝)과 Railway(철도): 교통사고 심각성과 유의미한 관계
- 관련 심각성: Severity 3으로 높은 수준 확인

Suggestion

- No_Exit와 Railway 변수에 중점
- 도로 끝과 철도 주변의 Severity 3 교통사고 방지 방안 마련

Possible Solution

- 도로 끝에서 발생한 사고: Street 이름 기준으로 개수 세기 및 내림차순 정렬
    - 결과:
        - Woodruff Rd (324건)
        - Brooklyn Queens Expy (261건)
        - I-95 N (258건)
        - SW 107th Ave (252건)
    - 제안: 해당 거리의 도로 연결 또는 교통 통행량 조절
- 철도 주변에서 발생한 사고: 같은 방법으로 탐색
    - 결과:
        - I-105 E (966건)
        - Glenn Anderson Fwy W (802건)
        - Glenn Anderson Fwy E (592건)
        - I-580 W (504건)
    - 제안: 철도 이외의 사고 영향 요인 분석 및 해당 거리 교통사고 방지 조치 마련