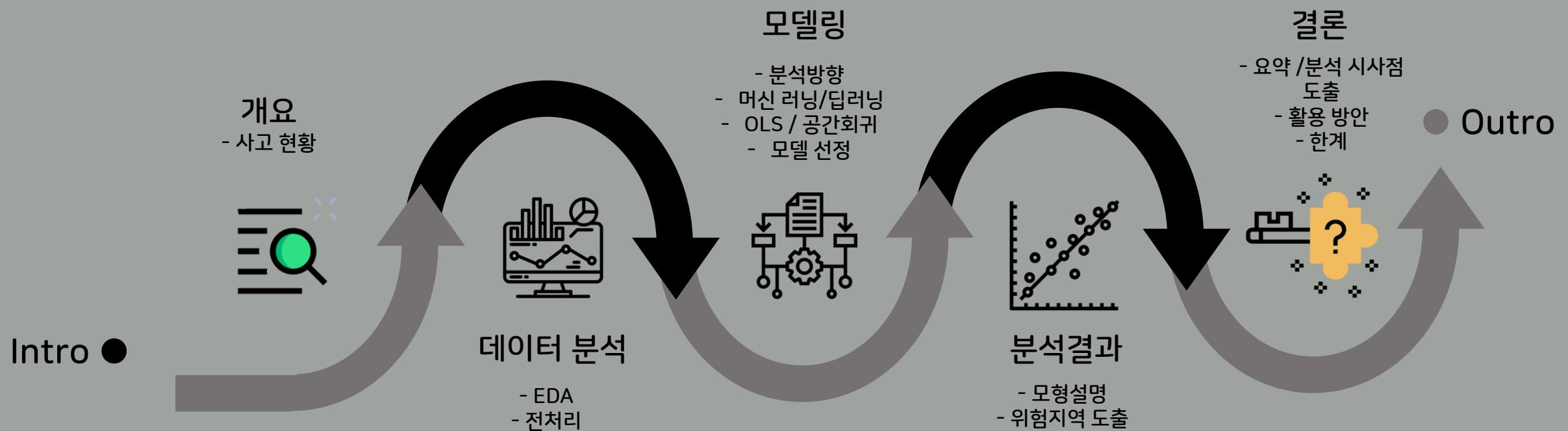


# 대전시 교통사고 위험지역 도출

Team Meetwin

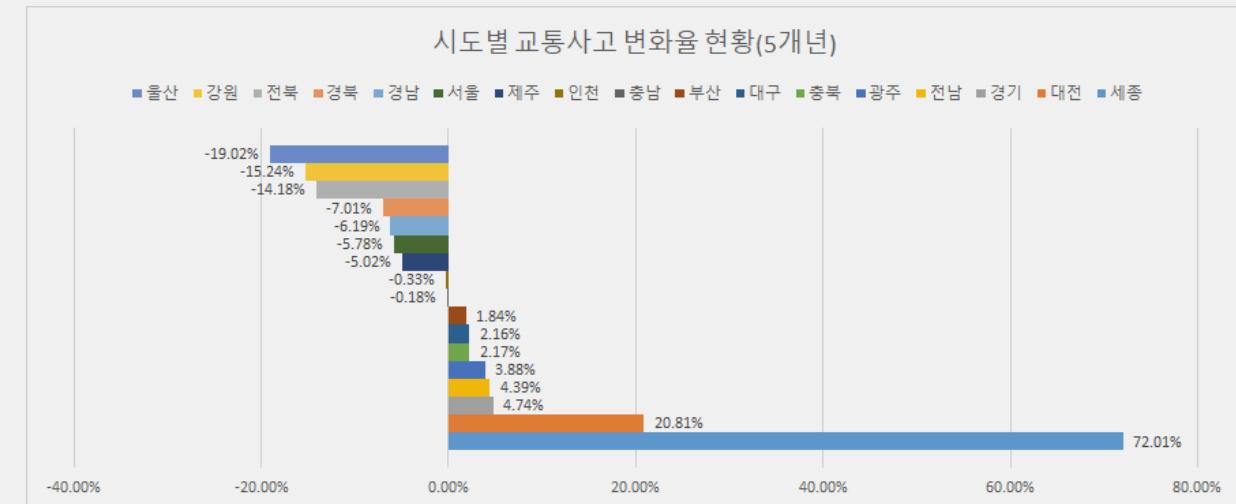
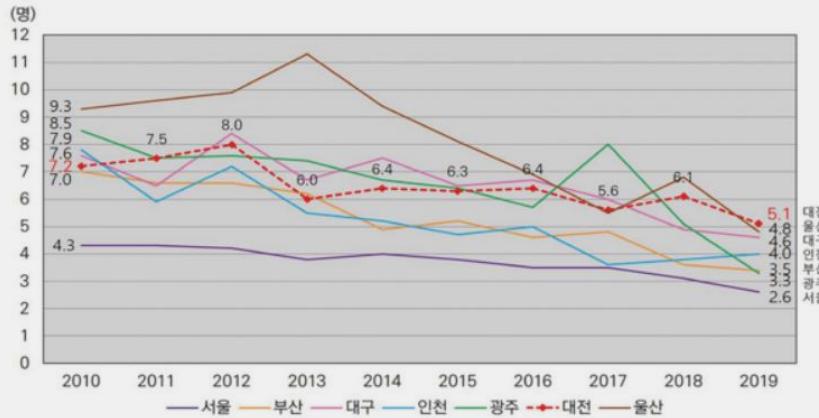


# INDEX



분석 개요

2010년 10만명당 7.2명→2019년 5.1명...감소율 29%  
광주(61.2%) 절반 수준



**대전**은 전국의 대도시 중 교통사고 **사망자 감소율**이 가장 낮고,  
**교통사고 증가율**은 20.81%로 세종에 이어 두번째로 높다.



시설개선사업 후 교통사고 사망자 및 사고건수 감소 효과/행정안전부 제공

#### ▶ 교통사고 잡은 곳 개선사업

- 교통사고 잡은 곳에 대한 사고요인 분석과 현장조사를 통해 도로구조, 교통시설 및 운영 측면의 문제점 개선

사업수행 실적	교통사고 잡은 곳 기본개선설계(1988~2020년): 21,644개소, 502개 구간		
---------	---	--	--

※ 개선 공사 지점 효과분석 결과('91~'20년)

구분	개선 전	개선 후	효과
발생건수(건)	204,902	146,261	28.6% 감소
사망자수(명)	3,997	2,208	44.8% 감소
부상자수(명)	196,271	140,399	28.5% 감소

교통시설물 설치와 도로 환경개선 등의 환경적 요인도 실제로 교통사고 감소에 영향을 끼치고 있다.

设施물과 도로 환경을 중심으로 교통사고를 감소시키는데 유의미한 요인이 무엇인지 분석할 필요가 있다.

# 데이터 분석



**EDA**



COMPAS(한국토지주택공사)에서 제공되는 데이터셋의 형태는  
**MultiPolygon, Points(좌표), LinePath(도로정보)**와 기타 공간 데이터로 나눌 수 있다.

### MultiPolygon

- 2. 대전광역시\_교통사고격자(2017~2019).geojson
- 5. 대전광역시\_안전지대.geojson
- 6. 대전광역시\_횡단보도.geojson
- 8. 대전광역시\_정차금지지대.geojson
- 12. 대전광역시\_인구정보(총인구).geojson
- 13. 대전광역시\_인구정보(고령).geojson
- 14. 대전광역시\_인구정보(생산가능).geojson
- 15. 대전광역시\_인구정보(유소년).geojson
- 23. 대전광역시\_도로명주소(건물).geojson
- 24. 대전광역시\_건물연면적\_격자.geojson
- 25. 대전광역시\_법정경계(시군구).geojson
- 26. 대전광역시\_법정경계(읍면동).geojson
- 27. 대전광역시\_행정경계(읍면동).geojson
- 28. 대전광역시\_연속지적도.geojson

### Point / LineString

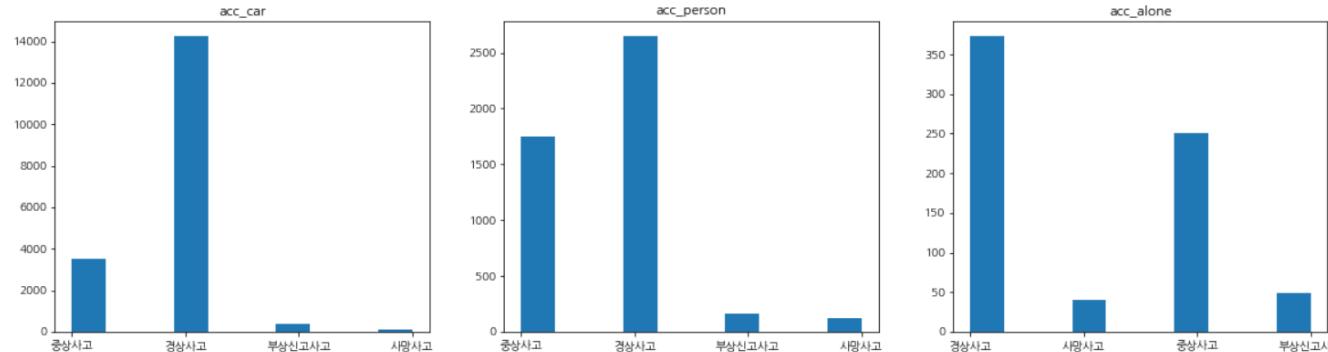
- 3. 대전광역시\_신호등(보행등).geojson
- 4. 대전광역시\_신호등(차량등).geojson
- 7. 대전광역시\_도로속도표시.geojson
- 9. 대전광역시\_교통안전표지.geojson
- 17. 대전광역시\_교통링크(2018).geojson
- 18. 대전광역시\_교통노드(2018).geojson
- 19. 대전광역시\_상세도로망(2018).geojson

### 기타 공간 데이터

- 1. 대전광역시\_교통사고내역(2017~2019).csv
- 11. 대전광역시\_동별\_인구현황(2019~2019).csv
- 16. 대전광역시\_기상데이터(2017~2019).csv
- 20. 대전광역시\_평일\_일별\_시간대별\_추정교통량(2018).csv
- 21. 대전광역시\_평일\_일별\_혼잡빈도강도(2018).csv
- 22. 대전광역시\_평일\_일별\_혼잡시간강도(2018).csv
- 29. 코드정의서.xlsx

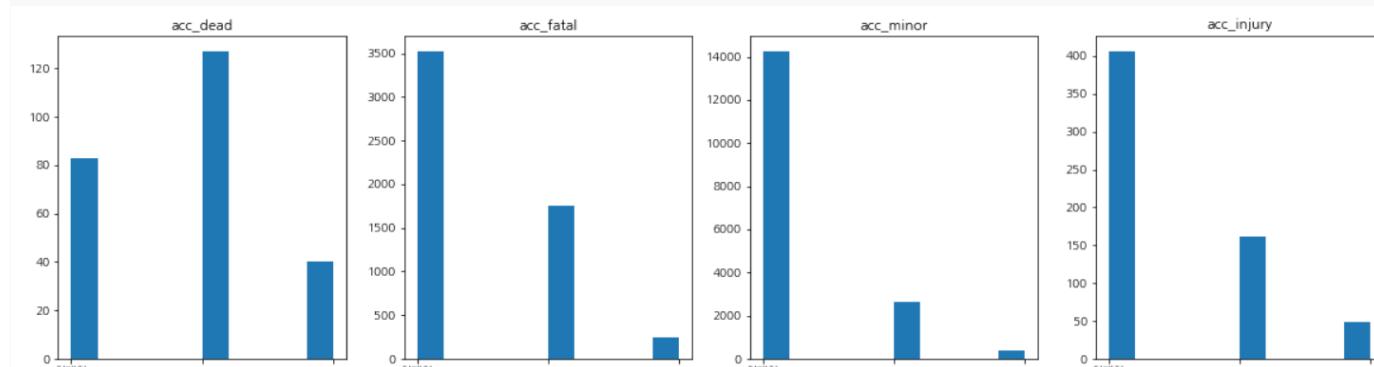
## 사고 유형별 특징 파악

: 전체 사고 기록 데이터를 사고 유형별로 세분화해서 각각의 특징을 분석



부록 [그래프1]

중상사고와 사망사고를 위험사고라 가정했을 때,  
각 유형별로 위험사고가 차지하는 비율은  
차량단독사고(57%) > 차대사람(39%) > 차대차(19.4%)이다.



반대로 각 사고 정도로 나누어 각 사고유형이 차지하는 비율을 봤을 때,  
사망자 발생 사고는 차대사람 사고 유형의 비율이 가장 높았으나,  
나머지 중상,경상,부상사고는 차대차 사고 유형의 비율이 높음을 알 수 있다.

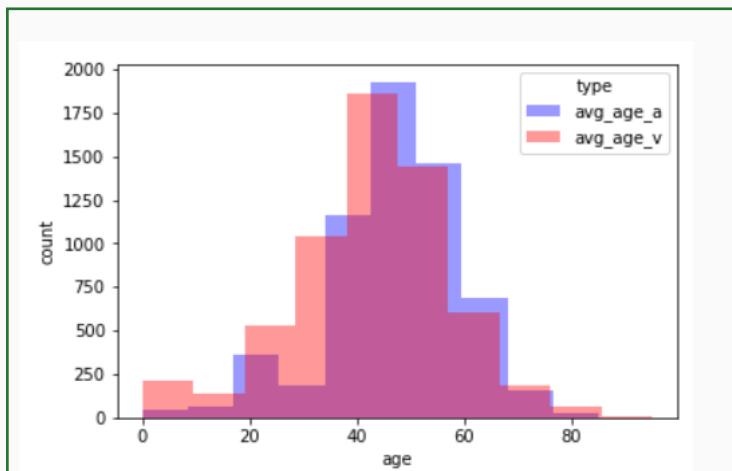
부록 [그래프2]



실제 사망자 수와 중상자 수에 많은 영향을 끼치는 사고유형은 차대사람, 차대차 사고 유형이라는 것을 알 수 있다.

연령대별 특징 파악

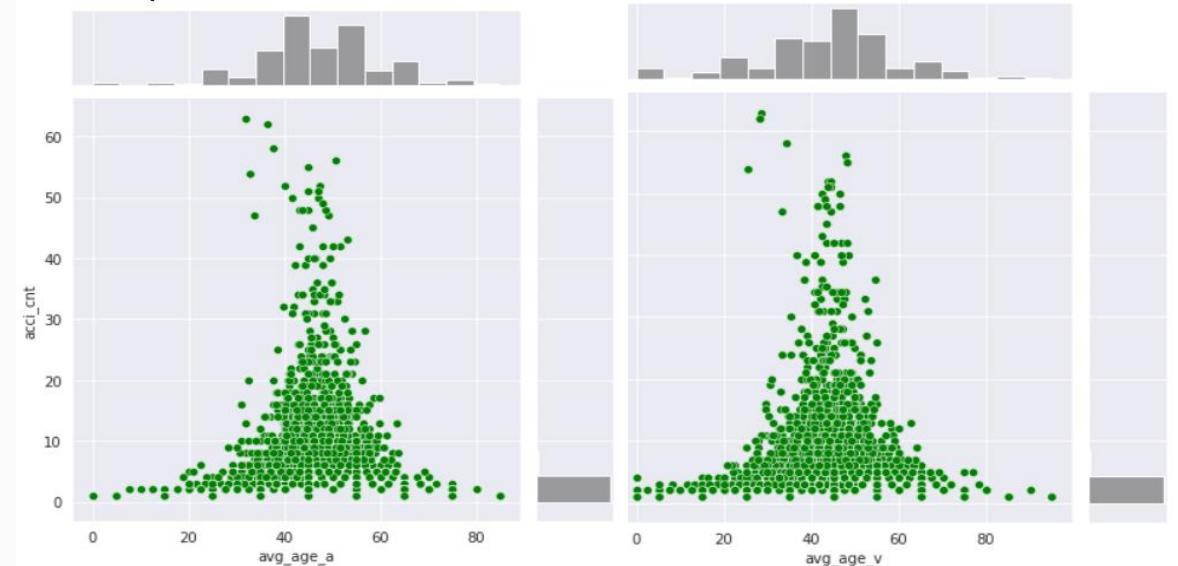
- 교통사고 격자별 가해자(age\_a)와 피해자(age\_v) 평균 연령 분포 시각화



부록 [그래프3]

두 분포를 봤을 때, 가해자와 피해자의 평균 연령 분포는 유사함  
다만, 피해자의 연령분포가 가해자에 비해 조금 더 어린 것으로 나타남

- 교통사고 격자별 가해자(age\_a)와 피해자(age\_v) 평균 연령 분포 시각화



부록 [그래프4]

사고건수와 가해자, 피해자의 평균 연령 분포는 선형관계가 보이지 않음

분석 결과, 특정 연령대가 특정 유형의 사고에 집중 되어있는 경향성을 발견하기 어려움

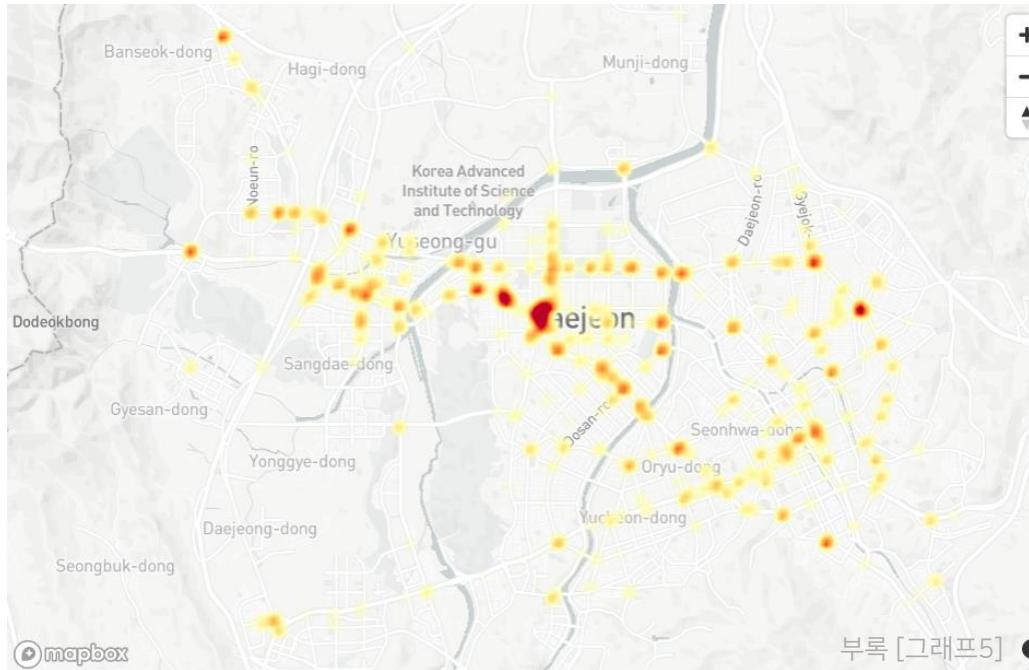
30-40대의 사고건수나 위험도가 높은 것은 해당 연령대의 인구분포가 높은 특성이 반영된 것으로 보임

따라서 교통사고와 발생과 가해자, 피해자의 연령대는 뚜렷한 연관성이 없어 보임



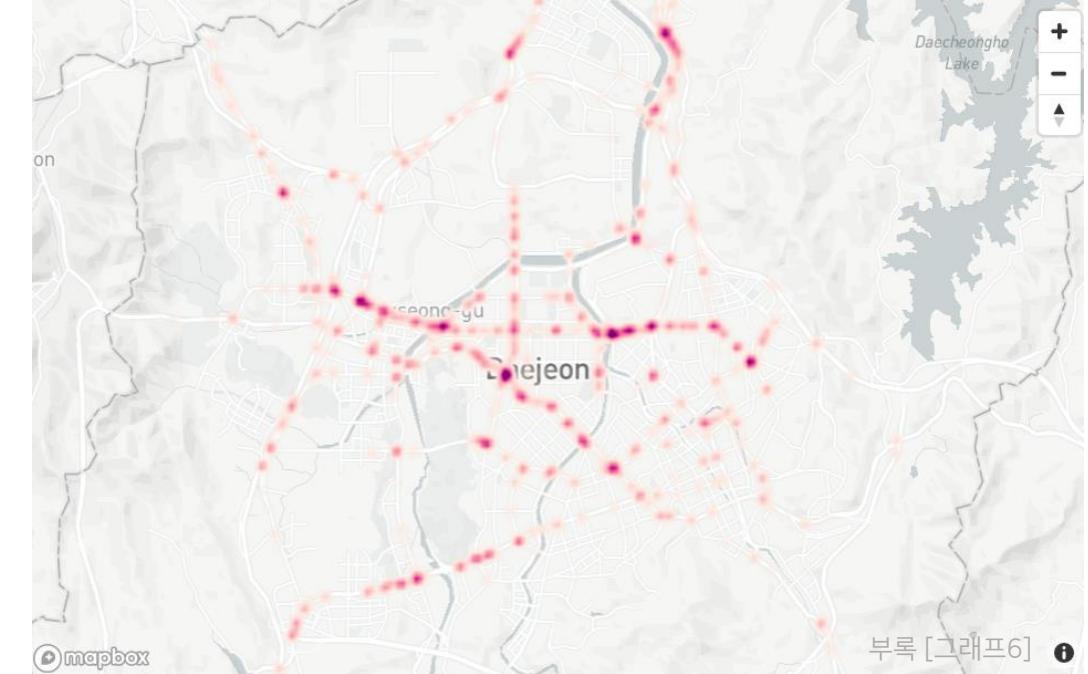
### 주요 변수 시각화

대전시 교통사고 데이터 시각화



도로를 따라 높은 밀도로 사고 발생, 유성구, 서구(계룡로)에 집중적으로 사고 발생

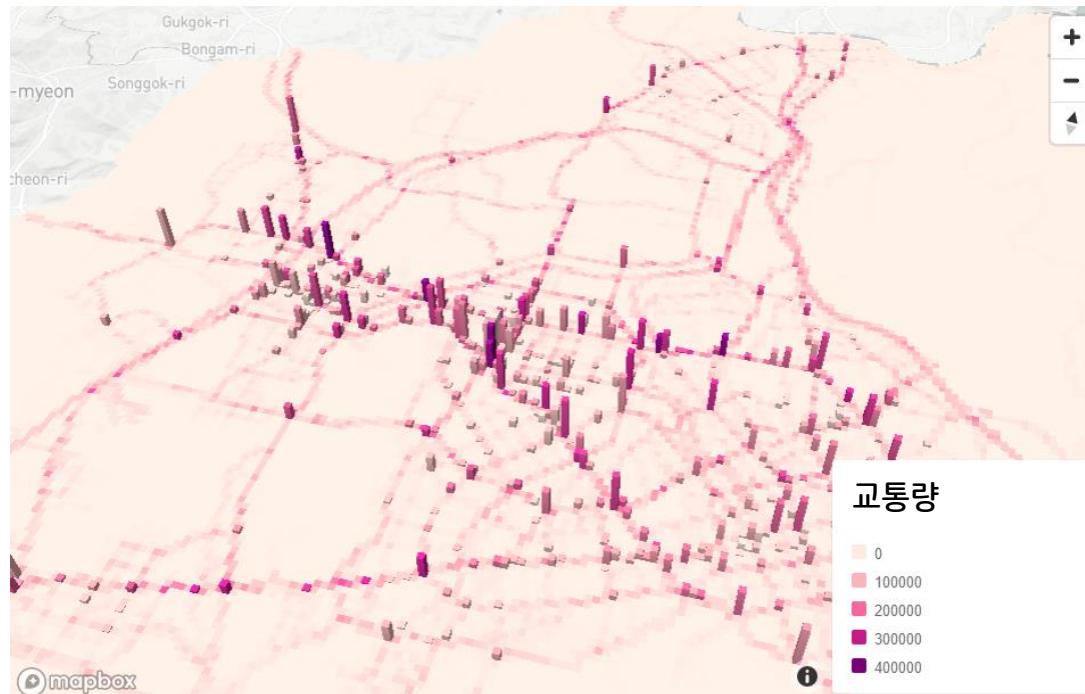
대전시 교통량 데이터 시각화



대전시청이 위치한 서구를 중심으로 높은 교통량이 분포

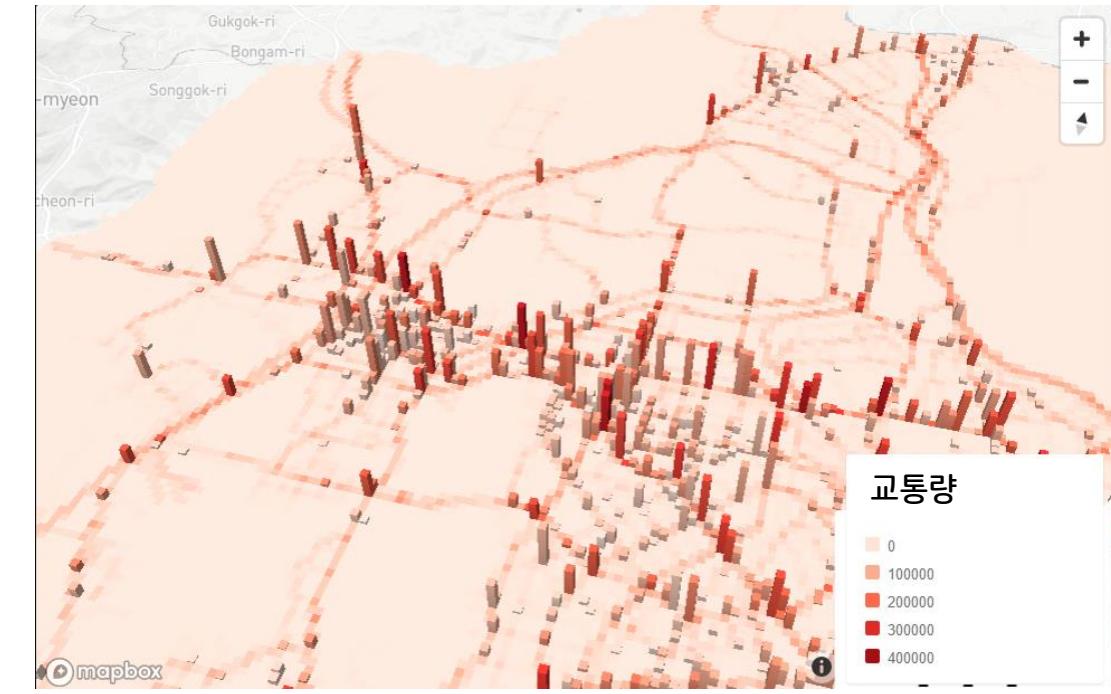
### 주요 변수 시각화

사고건수 + 교통량 데이터 오버레이



교통량과 사고 건수가 유사한 분포를 보임(사고건수: 높이)

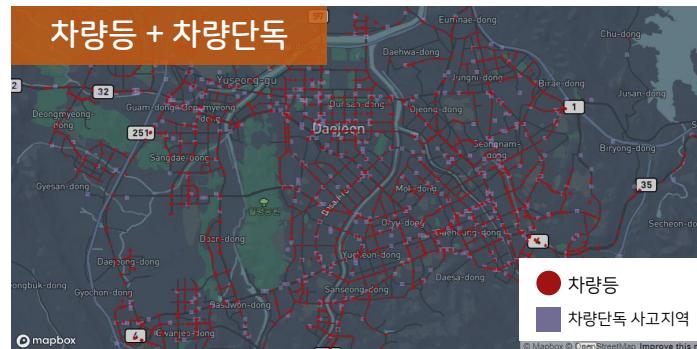
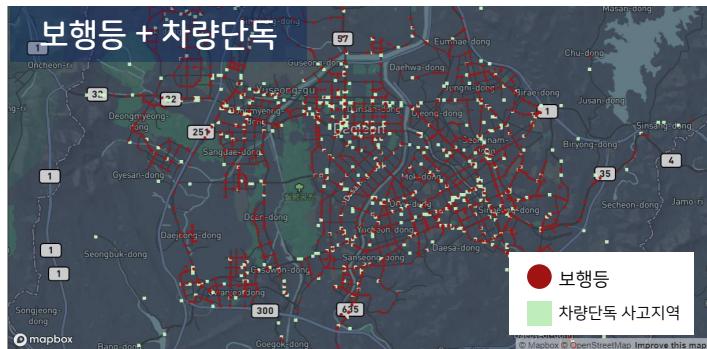
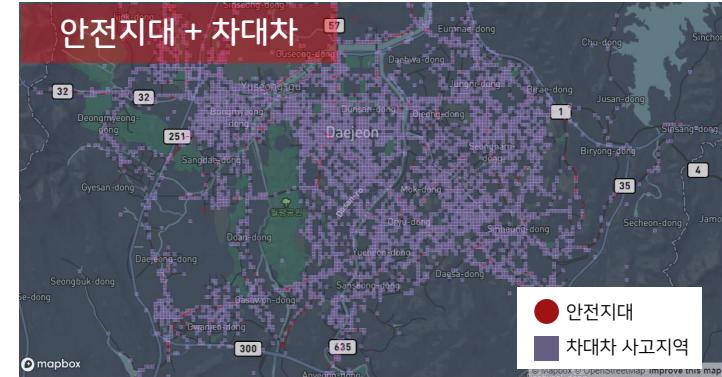
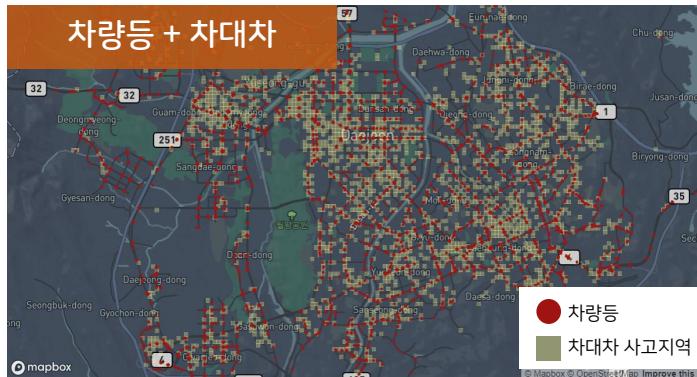
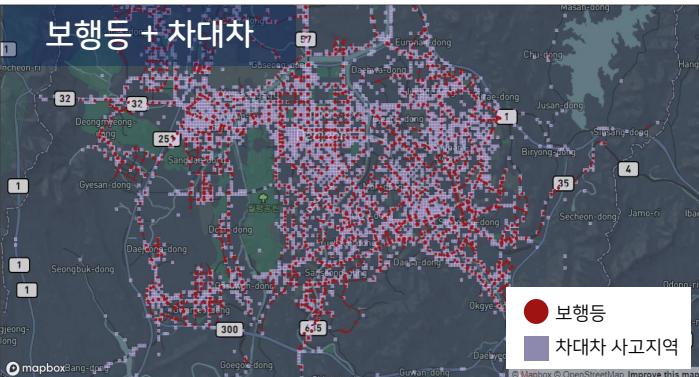
위험도(EPDO 지수) + 교통량 데이터 오버레이



교통량과 위험도 또한 유사한 분포를 보임  
사고건수보다 좀 더 뚜렷한 관계가 보임



**주요 변수 시각화** 사고건수와 밀접한 관련이 있어 보이는 변수들을 오버레이하여 시각화 함



사고 유형/상관계수	보행등	차량등	안전지대
차대차 사고	0.6	0.57	0.49
차대사람 사고	0.41	0.48	0.33
차량단독 사고	0.35	0.35	0.25

부록 [그래프9~16]



# 데이터 전처리



# 전처리

## : 종속변수 처리 (1)



사고 내역, 사고 격자 데이터에서 중복 및 누락된 데이터 처리

	유형
1	한 사고가 동일격자에 여러 번 집계된 사고데이터
2	사고 건수(사고 격자) 데이터에만 기록된 사고데이터
3	한 사고가 여러 격자에 중복 처리된 사고데이터



	처리 방법
1	1-1. 중복신고로 보이는 acc_hist.iloc[ [20208,20209] ] 중 하나 삭제 (taas 활용) 1-2. 사고격자 데이터에서의 사고 건수와 사고내역 데이터를 동일하게 변경
2	taas를 활용하여 사고 격자 정보값을 검색, 동일 사고 탐색
3	taas를 활용해 해당 사고의 정확한 위치좌표 검색 후, 좌표를 포함하는 격자에 사고 데이터 할당, 나머지 격자는 데이터 삭제



사고유형/사고정도 데이터 범주화

gid	acci_cnt(사고건수)
다바866110	0
.	.
.	.
.	.
다바931203	3



acci_cnt(사고건수)	차대차_사고건수	차대사람_사고건수	차량단독_사고건수	사망피해자 수	중상피해자 수	경상피해자 수	부상피해자 수
0	0	0	1	4	0	0	0
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
3	2	1	0	0	3	2	1

## EPDO(Equivalent Property Damage Only)

- 교통사고비용과 EPDO에 근거한 사고밀도 청주사례 분석, 박나영, 박병호(2018)

$$\text{EPDO 사고건수} = \text{사망사고건수} * 12 + \text{부상사고건수} * 3 + \text{계산피해만의 사고건수}$$

각 피해의 종류를 등가로 환산하여 하나의 피해단위로 나타내어 비교하는 지수

단순 사고건수가 아닌 위험도를 계산해 분석에 사용함으로서 사고의 경증까지 고려한 분석이 가능

acci_cnt(사고건수)	사망 피해자 수	중상 피해자 수	경상 피해자 수	부상 피해자 수
0	4	0	0	0
.	.	.	.	.
.	.	.	.	.
3	0	3	2	1

>>

EPDO 지수
48
.
.
12

## 사고율

$$\text{accident}_{ratio} = \frac{\text{accident}_{count}}{\text{traffic}_{total\_sum}}$$

절대적인 사고건수보다 해당 지역의 교통량에 비례하여 실질적인 위험율을 계산



### 도로 관련 데이터 전처리 및 통합

#### 0. 대전지적도와 상세도로망 데이터 병합

#### 1. 교통량 및 도로정보 / 혼잡빈도강도 / 혼잡시간강도

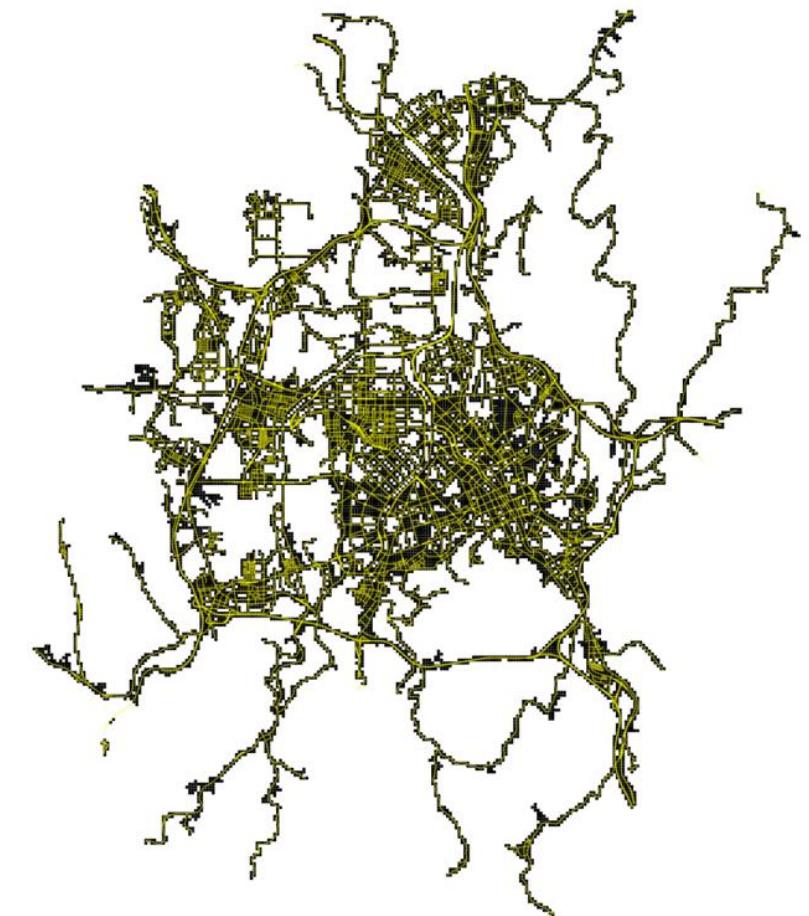
- 전일 시간대 데이터 선택
- Link\_id를 오산시 상세도로망의 Link\_id로 병합
- 격자에 데이터를 할당하기 위해 교통사고 격자와 linestring의 intersects를 적용하여 join
- 이 때, 사고는 발생했으나 교통량은 0인 격자가 발생(=교통량이 null인 경우)
- 대전광역시\_연속지적도'를 활용하여 추가적인 도로망 파악 및 null인 교통량 대체

#### 2. 교차로

- 교통노드 데이터의 NODE\_TYPE변수를 활용하여 교차로인 것들만 추출
- 교통사고 격자와 join



### 최종 교통량 할당 데이터



“

교통 사고 관련 데이터의 특성을 집중적으로 분석하기 위해 도로가 존재하는 지역에서 분석 진행

”

## 내부 데이터 전처리 - Point type

### 건물 관련 변수

#### 1. 4510개의 격자별로 포함하는 건물(전체 / 주거용 / 사무용) 개수 측정

building - 한 격자가 포함하는 건물의 개수

office - 코드정의서에 기입된 건물용도에 따라 사무용으로 분류된 건물의 개수

※건물용도코드가 04401~04499, 10000, 10101~10299에 해당하는 건물

house - 코드정의서에 기입된 건물용도에 따라 주거용으로 분류된 건물의 개수

※건물용도코드가 01000~02007에 해당하는 건물

#### 2. '대전시\_어린이교통사고\_격자' 데이터셋과 격자를 기준으로 하여 **join**함

### count 변수

#### 1. 신호등(보행등/차량등): 해당 격자가 포함하는 신호등의 개수 count

#### 2. CCTV: 해당 격자가 포함하는 CCTV의 개수 count

#### 3. 교통안전표지: 해당 격자가 포함하는 교통안전표지의 개수 count

### 면적 관련 변수

#### 1. 횡단보도: 해당 격자가 포함하는 횡단보도의 면적 계산

#### 2. 정차금지지대: 해당 격자가 포함하는 정차금지지대의 면적 계산

### 기타 변수

#### 1. 중앙분리대: 해당 격자가 중앙분리대를 포함하는 지의 여부

#### 2. 안전지대: 해당 격자가 안전지대를 포함하는 지의 여부

#### 3. 인구데이터 / 차량등록현황: 격자별로 인구 수/차량등록 수의 합계

### 외부데이터 수집

#### 사고지점 반경 주요시설 (카카오 지도/로컬 OPEN API 활용)

#### 격자의 중심좌표 반경 50M 내에 있는 주요 시설물들의 개수를 계산

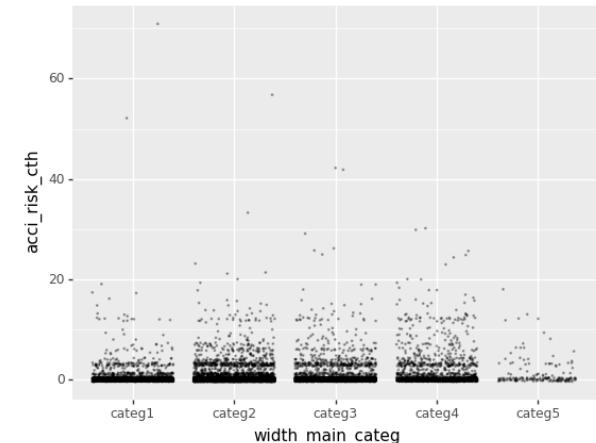
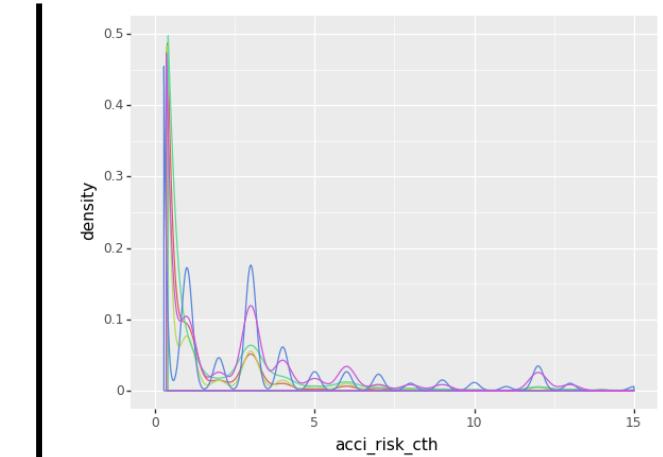
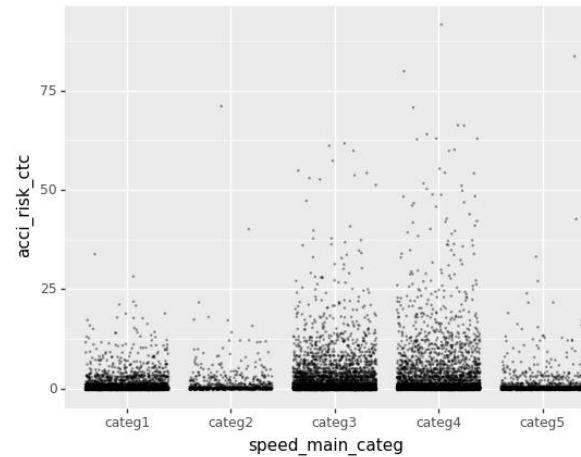
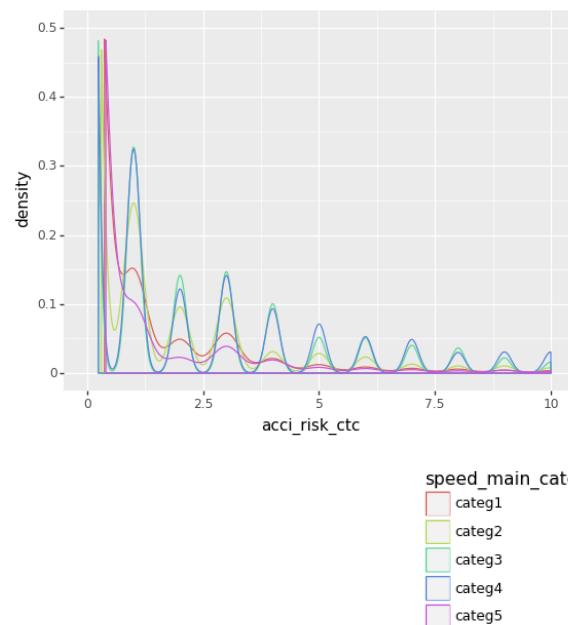
시설물 종류: 지하철 역, 주차장, 학교, 대형마트, 식당

 파생변수

'도로속도제한' 변수와 '차선' 변수를 활용해 '**격자의 주요 속도**', '**주요도로의 차선**'이라는 파생변수 도출

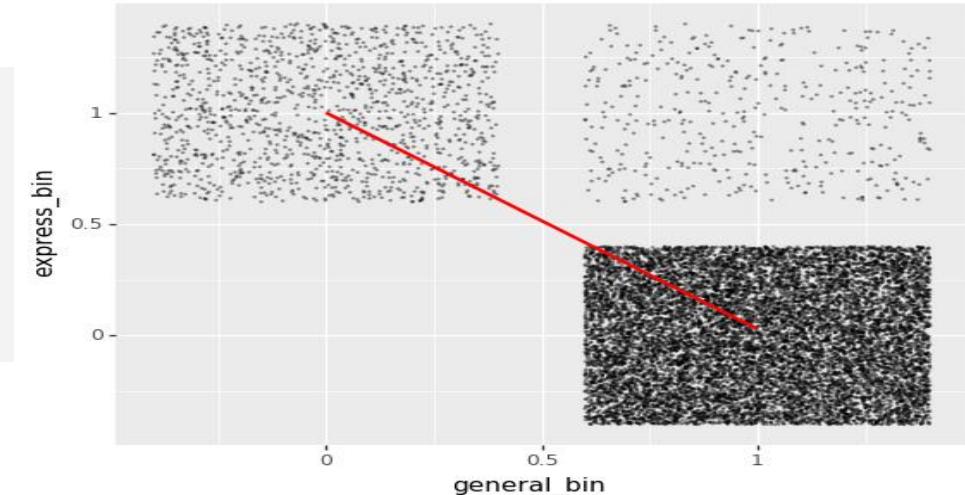
각 파생변수의 범주에 따른 사고 위험도의 분포가 다름을 통해

격자의 속도와 차선 정보가 사고 위험도에 유의미한 영향을 주는 것을 알 수 있음.

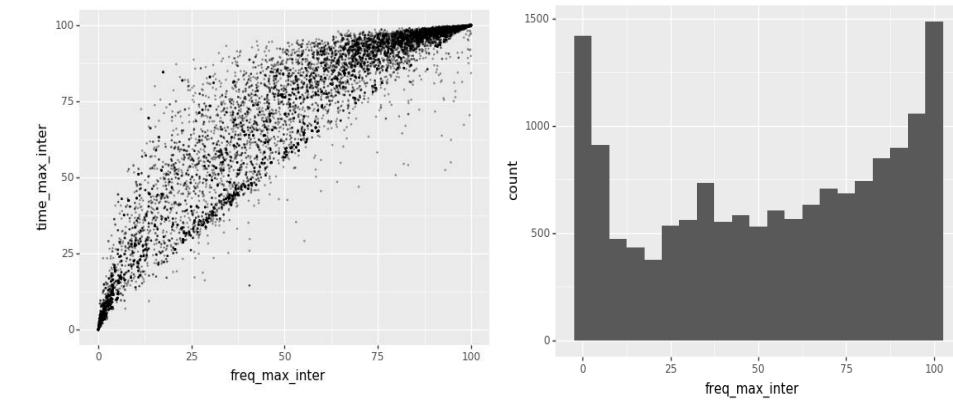


파생변수

express, general은 두 변수 사이의 선형성이 강하기 때문에  
두 변수를 하나의 변수로 통합하여 express 변수로 차원을 축소함



마찬가지로 freq\_max(혼잡빈도강도\_최대치), time\_max(혼잡시간강도\_최대치)도  
선형성이 강해 분포가 고른 freq\_max를 사용하기로 함



최종 전체 데이터셋

일부 중복되는 의미의 변수를 정제함

## 종속변수 관련 변수

- acci\_cnt
- acci\_risk
- acci\_ratio
- acci\_car\_car
- acci\_risk\_ctc
- acci\_ratio\_ctc
- acci\_car\_human
- acci\_risk\_cth
- acci\_ratio\_cth
- acci\_car\_only
- acci\_risk\_cal
- acci\_ratio\_cal

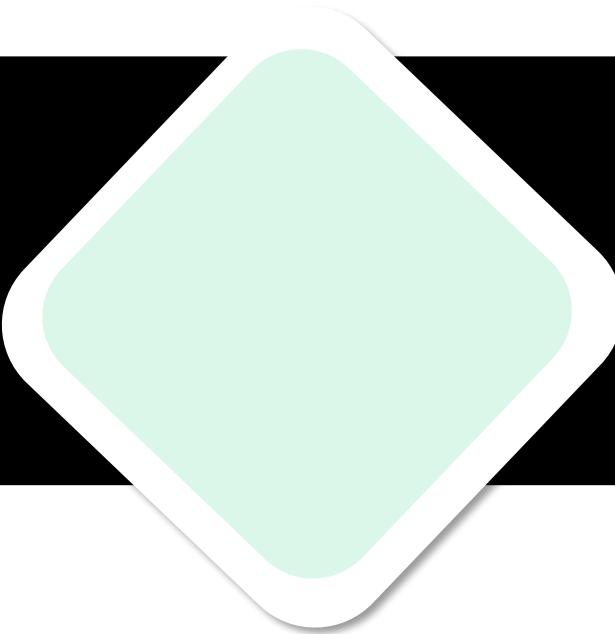
## 교통량 관련 변수

- traffic\_total\_sum\_inter
- freq\_max\_inter
- time\_max\_inter
- speed\_0
- speed\_30
- speed\_40\_to\_50
- speed\_60\_to\_70
- speed\_80\_to\_110
- speed\_main\_categ
- width\_main\_categ
- width\_1
- width\_2
- width\_3
- width\_4
- width\_5

## 사회,인구 관련 변수

- express
- general
- oneway
- car\_lane
- intersects
- TURN\_P
- crosswalk\_ar
- noparking\_ar
- trlight\_pas
- trlight\_car
- cctv
- sign\_traffic
- safe\_zone
- barrier\_road
- pop\_all
- pop\_sen
- pop\_prd
- pop\_chd
- cars\_cnt
- building\_tot
- house
- office
- other
- subway\_cnt
- parkinglot\_cnt
- school\_cnt
- ssm\_cnt
- restaurant\_cnt

모델링



**분석방향**



### Problem

교통사고에 영향을 끼치는 요인을 파악하고,  
높은 사고 발생건수가 예측되는 위험지역을 도출

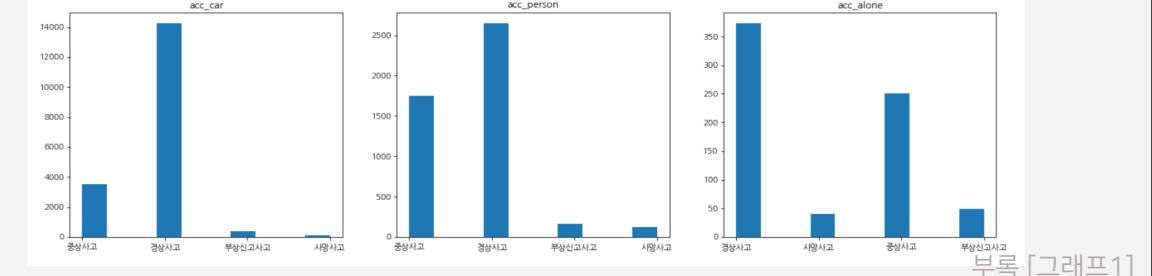


종속변수(사고건수)에 대해 각 요인이 가지는 영향력의 해석이 용이하며 모델 결과를  
토대로 위험지역을 예측할 수 있는 **회귀분석**으로 분석 방향을 설정

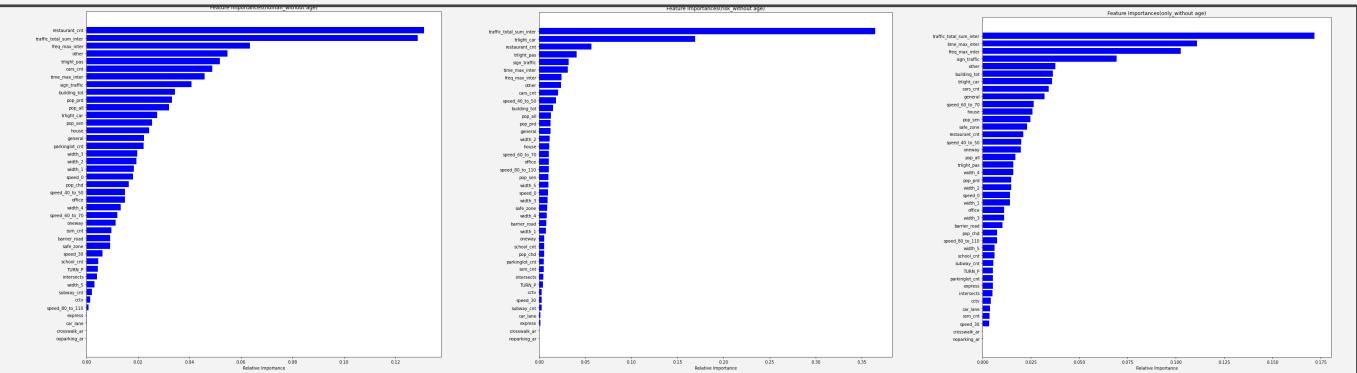
사고 유형별로 데이터를 나누어 회귀분석을 진행할 필요성 있음



1. EDA 결과 각 사고유형별로 위험사고(사망,중상사고)의 비율이 다르고  
각 사고 발생 건수의 큰 차이가 있음



2. RandomForest Feature importance를 분석한 결과  
사고유형별로 중요한 변수의 구성이 상이함



사고 유형별로 데이터셋을 나누어 분석하는 방식으로 진행

모형 탐색

OLS

머신러닝

딥러닝



변수 선택



공간회귀분석(Spatial Regression)

SEM

SLM

GWR



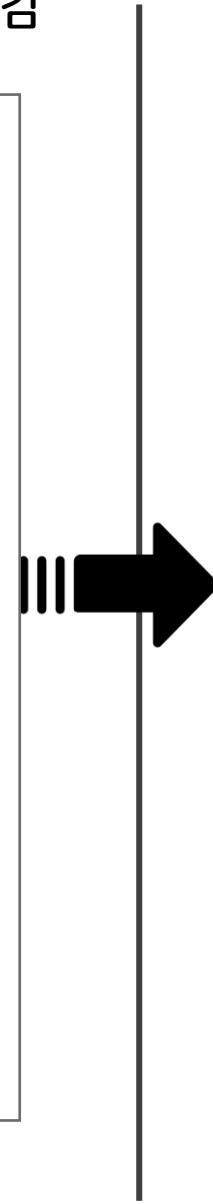
머신러닝 / 딥러닝



최종적으로 생성된 데이터셋으로 다양한 모델들을 학습시킴

반응변수 (종속변수)
'사고횟수_epdo'
'사고횟수_ratio'

설명변수 (독립변수)
'traffic_total_sum_inter', 'oneway', 'car_lane', 'barrier_road', 'speed_0', 'speed_30', 'express', 'general', 'width_1', 'width_2', 'width_3', 'width_4', 'freq_max_inter', 'intersects', 'TURN_P', 'building_tot', 'house', 'office', 'other', 'crosswalk_ar', 'noparking_ar', 'trlight_pas', 'trlight_car', 'cctv', 'sign_traffic', 'safe_zone', 'pop_all', 'pop_sen', 'pop_prd', 'pop_chd', 'cars_cnt', 'subway_cnt', 'parkinglot_cnt', 'school_cnt', :ssm_cnt', 'restaurant_cnt', 'speed_40_to_50', 'speed_60_to_70', 'speed_80_to_110'



## 일차적으로 별도의 파라미터 튜닝 없이 모델의 성능 테스트

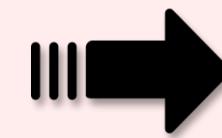
: 과적합을 피하고 모델의 성능 일반화를 위해 k cross validation 검정을 실시함

Algorithm	RMSE	MAE
SVR	4.201	1.541
Ridge regression	3.297	1.607
KNN regression	3.514	1.396
RF regression	3.144	1.341
LGBM	3.126	1.308
XGBM	3.279	1.393
LASSO regression	3.352	1.63
MLP	263.795	10.928

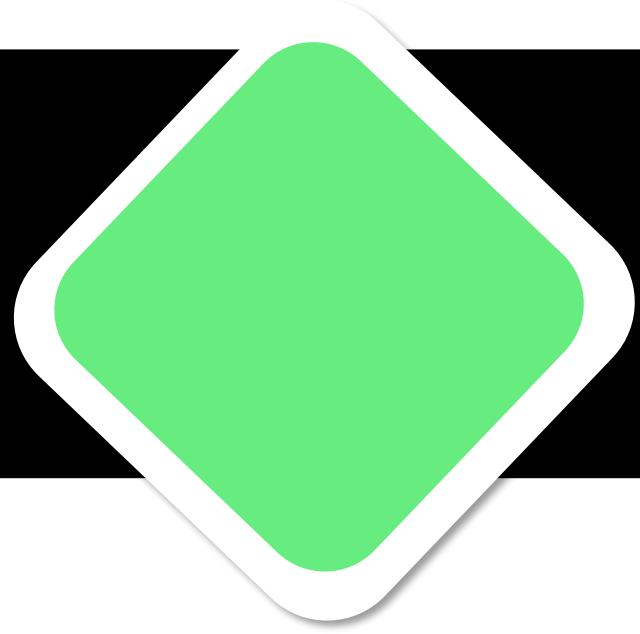
딥러닝 모델의 경우, 과적합에 민감한 특성이 있기 때문에 매우 좋지 않은 결과가 산출됨

머신러닝 모델 사이에서는 LGBM의 성능이 가장 좋지만 다른 모델들과 큰 차이는 나타나지 않음

딥러닝과 머신러닝 모델 모두 눈에 띄게 좋은 예측력을 보여주지 못했고,  
결과에 대한 해석이 어려워 교통사고에 영향을 주는 요인을 파악해야 하는 분석의 방향과 맞지 않다고 판단함



OLS 진행



**OLS / 공간회귀**



Ordinary Least Squares(OLS)

$$y = XB + \epsilon, \epsilon \sim N(0, \sigma^2)$$

오차항의 제곱을 최소로 만드는 회귀계수 추정법

하나로 합쳐져 있는 종속변수(사고건수\_EPDO지수, 사고율)를  
사고 유형별로 분리하여 3개의 모형 설정

DATA



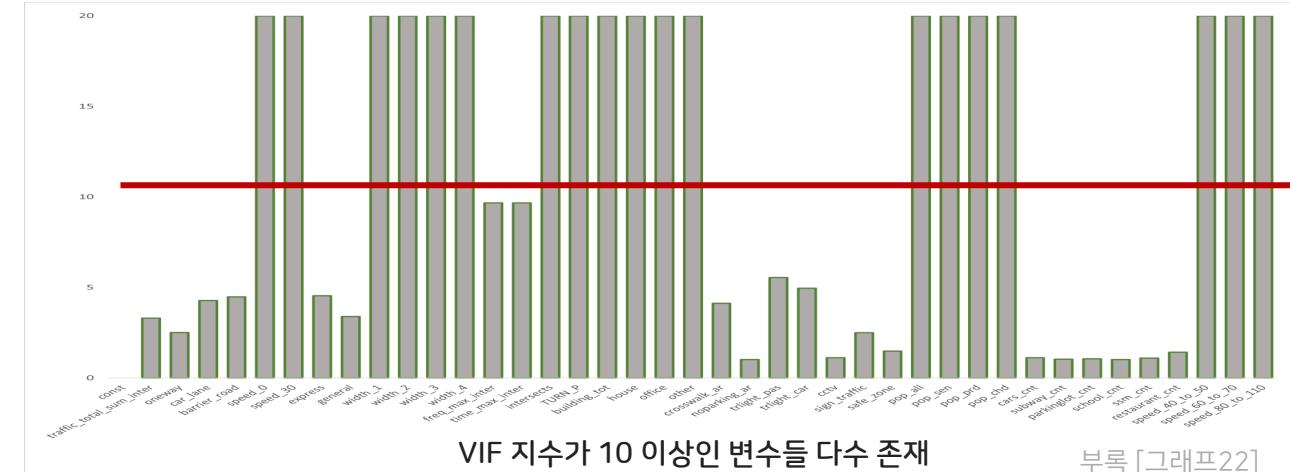
차대차

차대사람

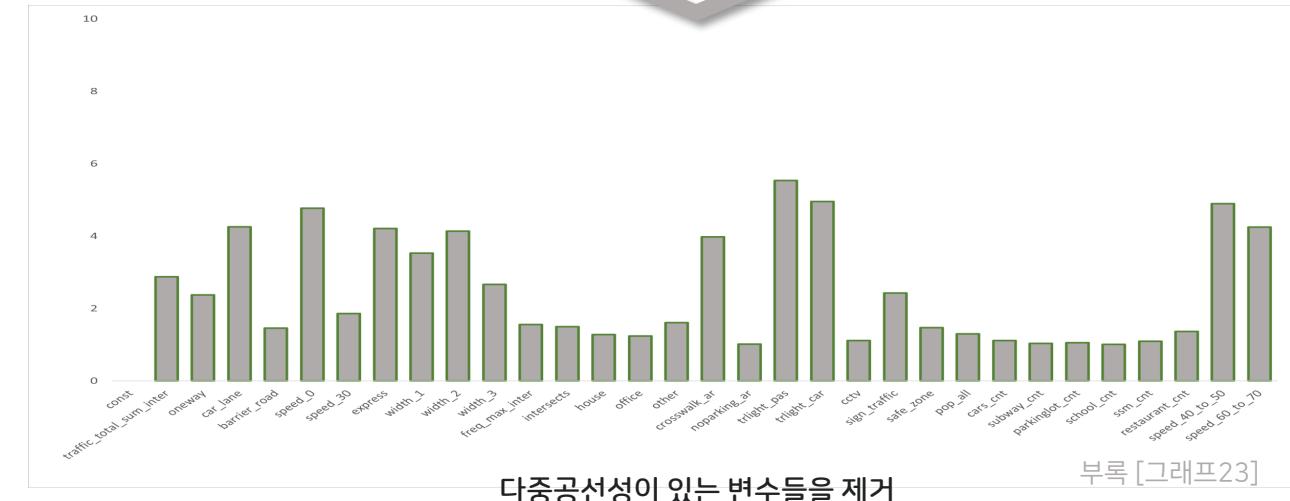
차량단독

각 모형별로 개별 분석 실시

다중공선성 확인



부록 [그래프22]



부록 [그래프23]

- 변수를 대상으로 Scaling 진행

*MinMax Scaler* 최대/최소값이 각각 1, 0이 되도록 스케일링

스케일링을 통해 다차원의 값들을 비교 분석하기 쉽게 만들어주며,  
자료의 오버플로우(overflow)나 언더플로우(underflow)를 방지

- 단계적 변수선택법 (Stepwise Regression) 진행 각 모형별로 선택된 변수의 목록

#### 차대차 모형

```
'traffic_total_sum_inter', 'crosswalk_ar', 'car_lane', 'cctv',
'restaurant_cnt', 'subway_cnt', 'trlight_car', 'safe_zone',
'office', 'barrier_road', 'oneway', 'speed_40_to_50', 'width_1',
'intersects', 'width_2', 'trlight_pas', 'other', 'express',
'freq_max_inter', 'pop_all', 'speed_30', 'noparking_ar', 'const'
```

#### 차대사람 모형

```
'resturant_cnt', 'crosswalk_ar', 'other', 'traffic_total_sum_inter',
'pop_all', 'car_lane', 'safe_zone', 'trlight_car',
'ssm_cnt', 'parkinglot_cnt', 'speed_30', 'office', 'subway_cnt',
'oneway', 'cctv', 'noparking_ar', 'house', 'speed_40_to_50', 'const'
```

#### 차량단독 모형

```
'traffic_total_sum_inter', 'trlight_car', 'cctv', 'other', 'subway_cnt',
'express', 'noparking_ar', 'school_cnt', 'width_2', 'const'
```

- 등분산성 검정(White test)

$$H_0 : \sigma_i^2 = \sigma^2 \text{ vs } H_1 : \sigma_i^2 \neq \sigma^2$$

TEST	DF	VALUE	p-value
White	264	5742.825	0.0000

세 모형 모두 등분산성을 만족하지 않음

- 공간자기상관성 : 공간상의 한 위치에서 발생하는 사건은 그 주변지역에서  
발생하는 사건과 높은 상관관계를 가짐(Spatial autocorrelation)

$$H_0 : I = 0 \text{ vs } H_1 : I \neq 0$$

$$I = \frac{N \sum_{i=1}^N \sum_{j=1}^N w_{ij} (Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{i=1}^N \sum_{j=1}^N w_{ij}) \sum_{i=1}^N (Y_i - \bar{Y})^2}$$

TEST	MI	VALUE	p-value
Moran's I	0.0989	20.440	0.0000

세 모형 모두 공간자기상관성이 있음



OLS는 등분산성을 만족하지 않고 공간자기상관성을 고려하기에 적합치 않음

단순한 OLS모형으로는 공간적 특성을 지닌 데이터를 설명하기에 충분하지 못하며  
왜곡된 추정결과를 낳으므로 이를 통제하기 위해 공간회귀모형을 사용



공간시차모형  
(spatial lag model; SLM)

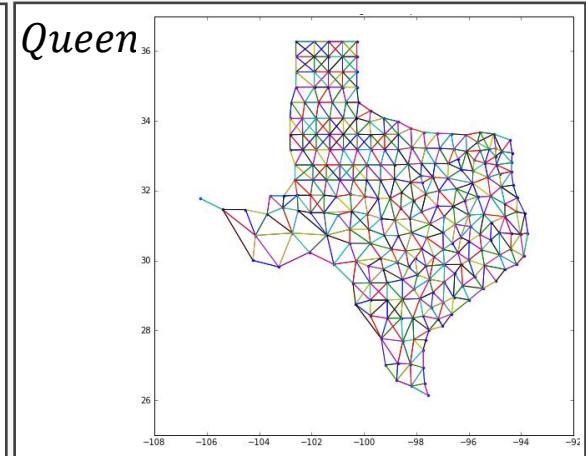
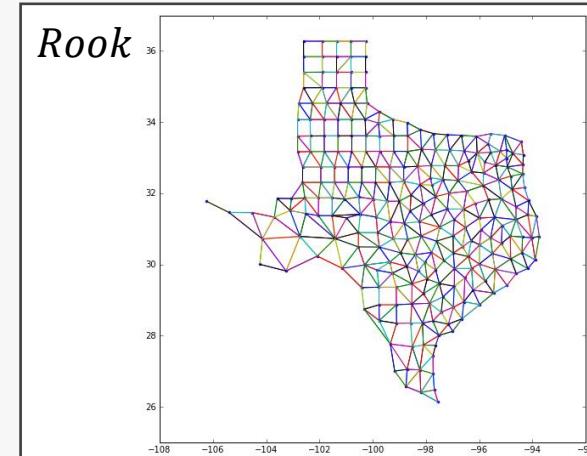
공간오차모형  
(spatial error model; SEM)

“ 위와 같은 공간회귀모형을 사용해 OLS 모형의 한계를 극복하고,  
이를 토대로 독립변수가 어린이 교통사고에 미치는 영향을 분석하기로 함 ”



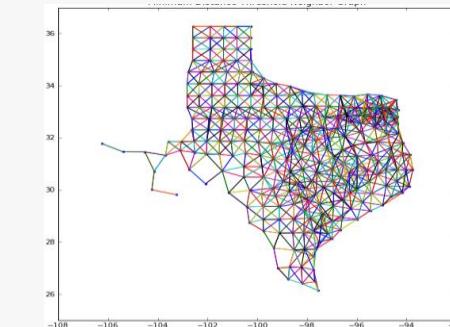
공간자기상관성을 확인하기 위해 **공간가중행렬**을 정의해야함

#### 1) 인접성척도를 기준으로 정의하는 방식



#### 2) 거리척도를 기준으로 정의하는 방식(DistanceBand)

- 500m를 기준으로 거리의 역수로 가중치를 부여한 행렬 정의



**Queen, DistanceBand** 방식으로 정의  
두 가지 방식 모두 행의 합이  
1이 되도록 정규화

# 모델링

## : OLS / 공간회귀 (4)

### LM Test / Robust LM 통계량 확인

#### Lagrange Multiplier, LM

LM test는 회귀모형에서 누락된 변수가 있는지를 검정하기 위해 적용됨.  
두 모형을 비교할 때 제약이 반영된 모형을 귀무가설로,  
제약이 없는 모형을 대립가설로 설정하여 비교함

#### Robust LM

Robust LM은 LM 통계치보다 적절한 공간회귀모형을 설정하는데 있어  
더 우수한 성능을 보임

### 차대차 사고\_모형 $\rightarrow$ 공간오차모형(SEM) 적합

공간가중행렬	TEST	DF	VALUE	p-value
Queen	Lagrange Multiplier(lag)	1	319.000	0.0000
	Lagrange Multiplier(error)	1	412.423	0.0000
	Robust LM(lag)	1	12.923	0.0003
	Robust LM(error)	1	106.347	0.0000
Distance Band	Lagrange Multiplier(lag)	1	660.582	0.0000
	Lagrange Multiplier(error)	1	1100.377	0.0000
	Robust LM(lag)	1	57.913	0.0000
	Robust LM(error)	1	497.708	0.0000

### 차대사람 사고\_모형 $\rightarrow$ 공간시차모형(SLM) 적합

공간가중행렬	TEST	DF	VALUE	p-value
Queen	Lagrange Multiplier(lag)	1	789.916	0.0000
	Lagrange Multiplier(error)	1	132.622	0.0000
	Robust LM(lag)	1	658.923	0.0000
	Robust LM(error)	1	1.628	0.2019
Distance Band	Lagrange Multiplier(lag)	1	1580.641	0.0000
	Lagrange Multiplier(error)	1	208.322	0.0000
	Robust LM(lag)	1	1548.987	0.0000
	Robust LM(error)	1	176.668	0.0000

### 차량단독 사고\_모형 $\rightarrow$ 공간시차모형(SLM) 적합

공간가중행렬	TEST	DF	VALUE	p-value
Queen	Lagrange Multiplier(lag)	1	9.853	0.0017
	Lagrange Multiplier(error)	1	4.868	0.0274
	Robust LM(lag)	1	7.932	0.0049
	Robust LM(error)	1	2.947	0.0861
Distance Band	Lagrange Multiplier(lag)	1	13.147	0.0003
	Lagrange Multiplier(error)	1	3.867	0.0492
	Robust LM(lag)	1	10.214	0.0014
	Robust LM(error)	1	0.934	0.3339

회귀계수 추정 방식 선택

공간오차모형  
(spatial error model:SEM)

ML

: 오차의 정규성을 만족하지 못함

공간시차모형  
(spatial lag model:SLM)

ML

: 오차의 정규성을 만족하지 못함

2SLS

: 정규성 가정 필요하지 않음

회귀계수를 추정하는 방식인 GM방식과 ML방식을 선택하기 위해 정규성 검정(Shapiro-wilk test)을 실시

$H_0$  : 표본의 모집단이 정규분포를 따른다 vs  $H_1$  : 따르지 않는다

$$W = \frac{(\sum_{i=1}^n a_i x_i)^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

모형	W	p-value
차대차_모형	0.369	0.0
차대사람_모형	0.292	0.0
차량단독_모형	0.118	0.0

 유형별 모형

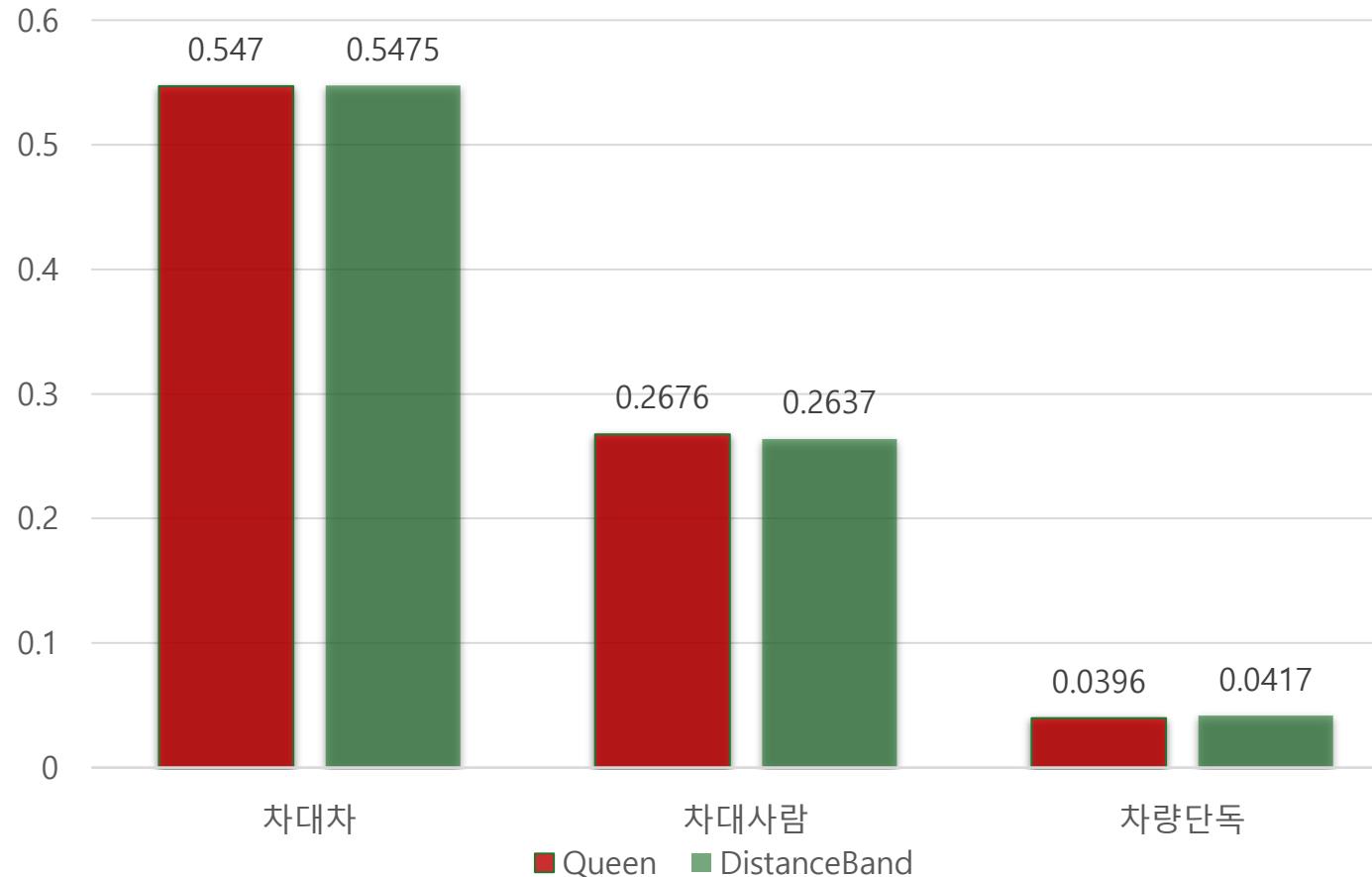
차대차: SEM - GM

차대사람: SLM - 2SLS

차량단독: SLM - 2SLS



성능 비교  
: R-squared



차대차를 제외한 나머지 모형의  
설명력이 충분하지 않음

모형의 성능 개선 필요

지리적가중회귀모델(GWR, Geographically Weighted Regression)

지역별로 서로 다른 회귀모형을 추정하는 모형

$$Y_i = \beta_{0i} + \sum_{k=1}^m \beta_{ki} x_{ki} + \epsilon_i$$

대역폭 (bandwidth) 설정 : 대역폭은 i지역과 j지역 간의 거리  $D_{ij}$ 에 따른 가중치의 민감도를 나타냄

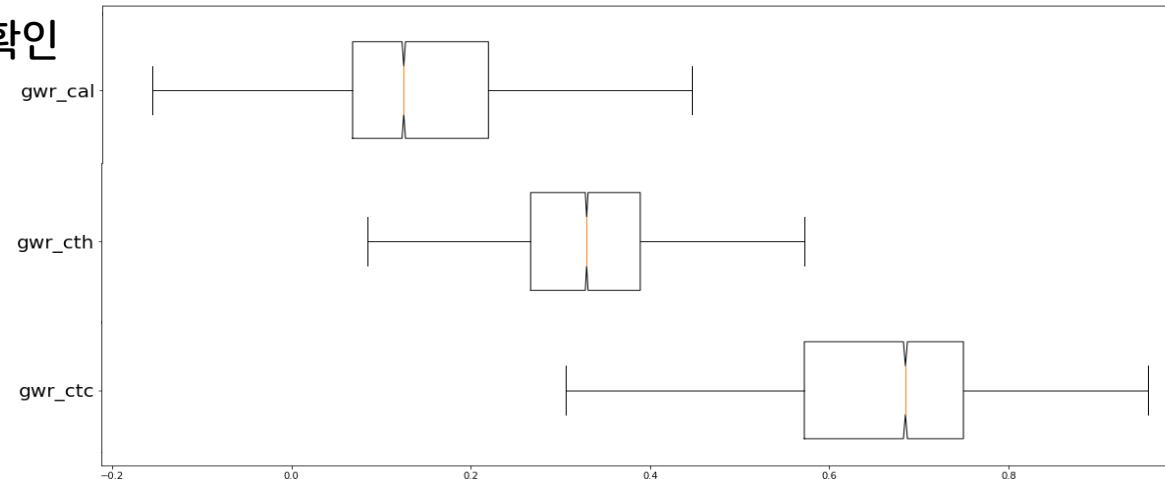
'가우시안' 커널을 사용해 최적의 대역폭 선정 >>

차대차: 89

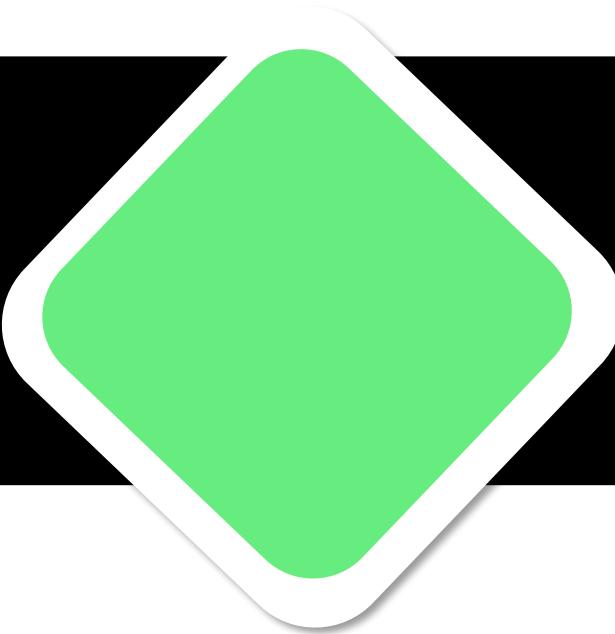
차대사람: 81

차량단독: 63

Local R-squared 분포 확인



SEM/SLM	GWR
0.5475	0.727
0.2052	0.399
0.0417	0.142

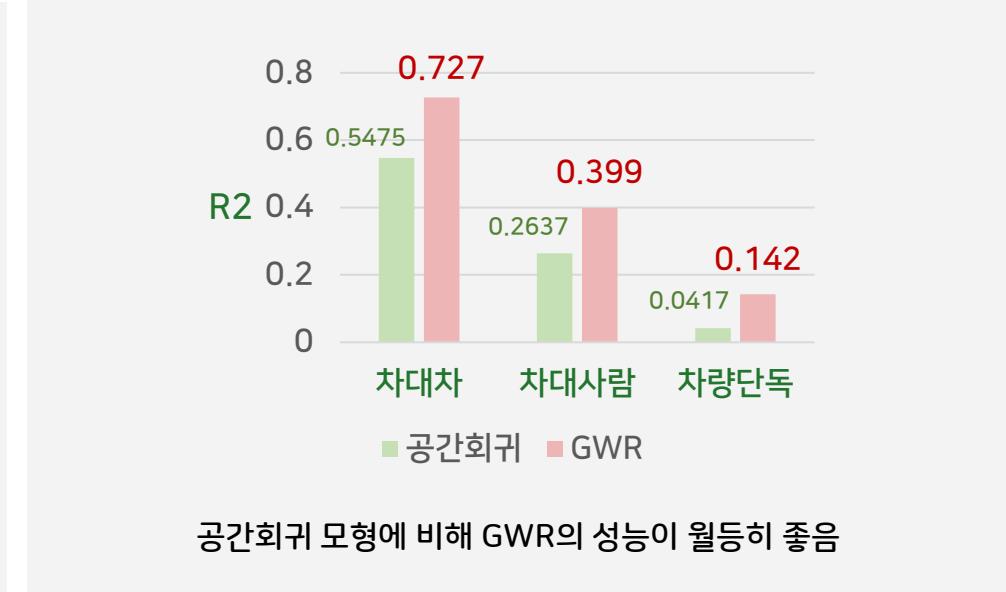
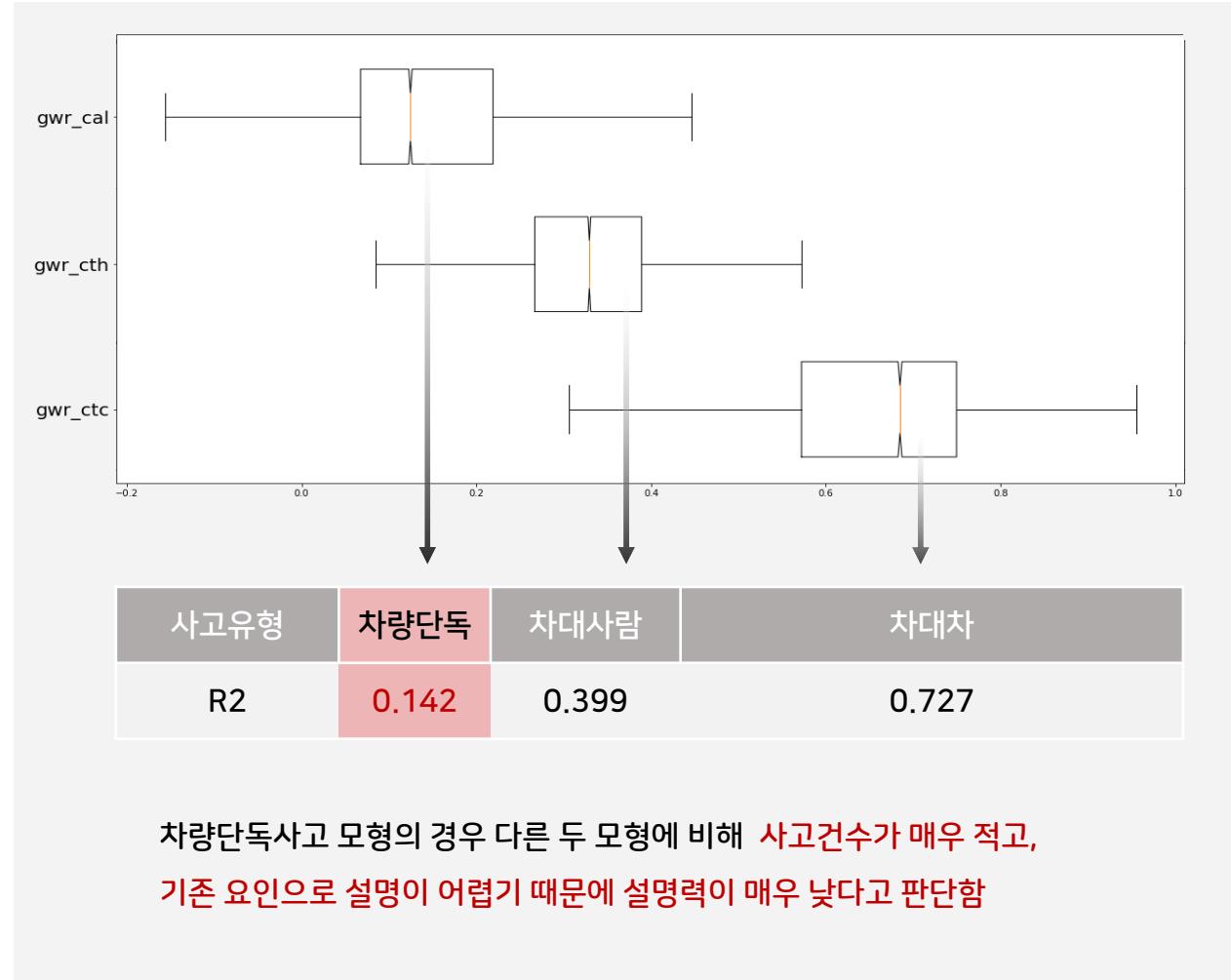


# 모델선정



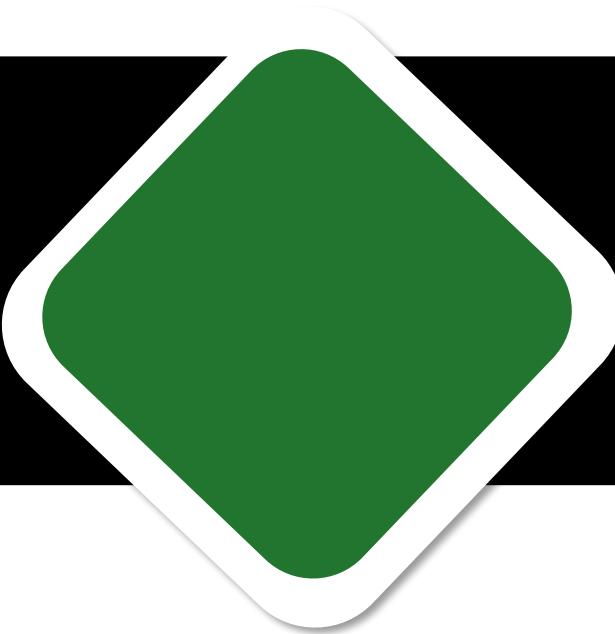
## 최종 모형 선정

모형의 성능이 가장 우수했던 **GWR(지리적가중회귀모델)**로 최종 모형 선정



GWR 모형 중 차대차, 차대사람사고  
두 개의 모형을 최종 모형으로 선정

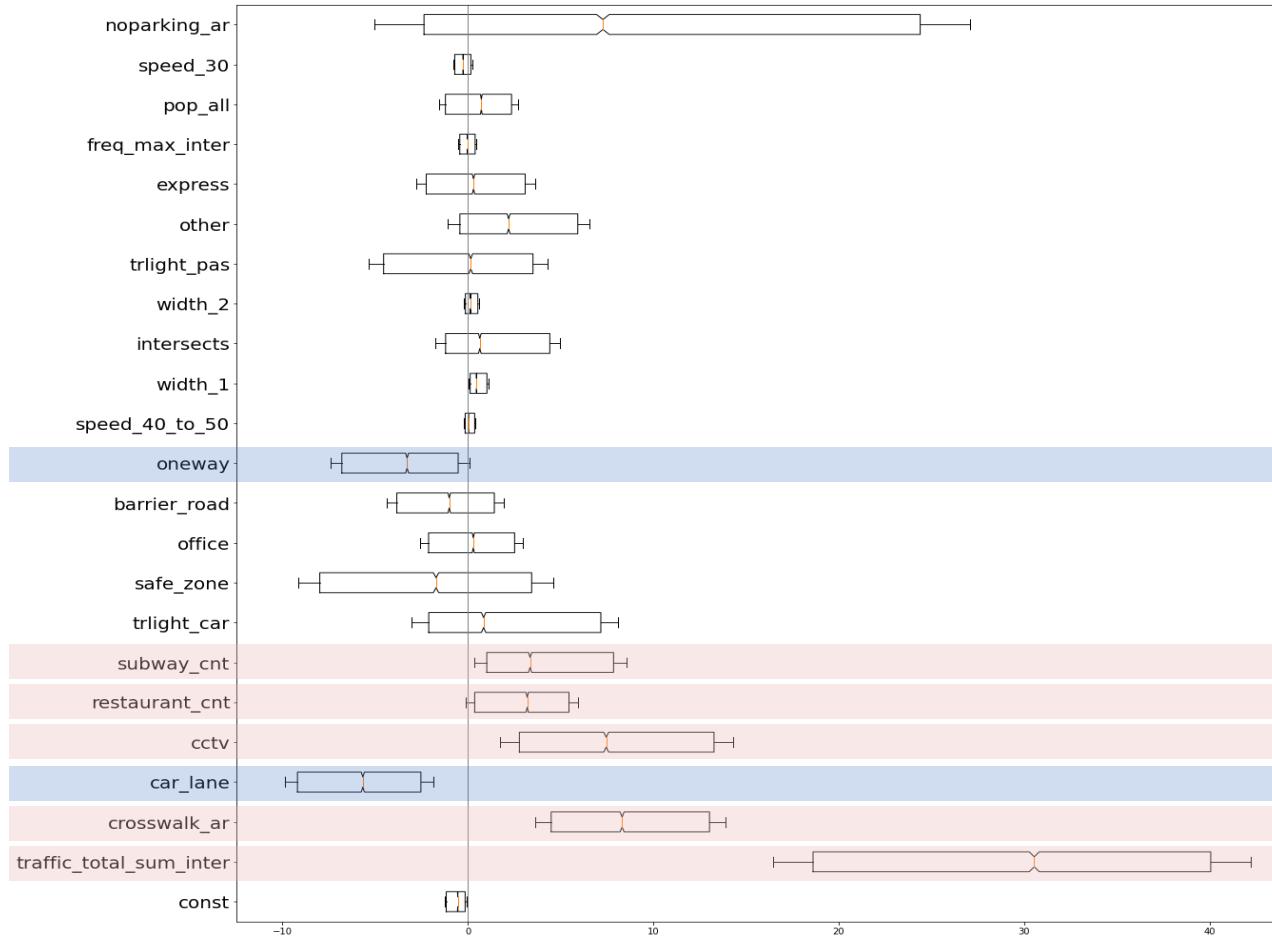
# 분석결과



# 모형 설명



## 차대차\_모형



영향을 많이 주는 요인

양의 관계를 가지는 요인

- 지하철(subway\_cnt)
- 음식점(restaurant\_cnt)
- 교통단속 cctv(cctv)
- 횡단보도 면적(crosswalk\_ar)
- 교통량  
(traffic\_total\_sum\_inter)

음의 관계를 가지는 요인

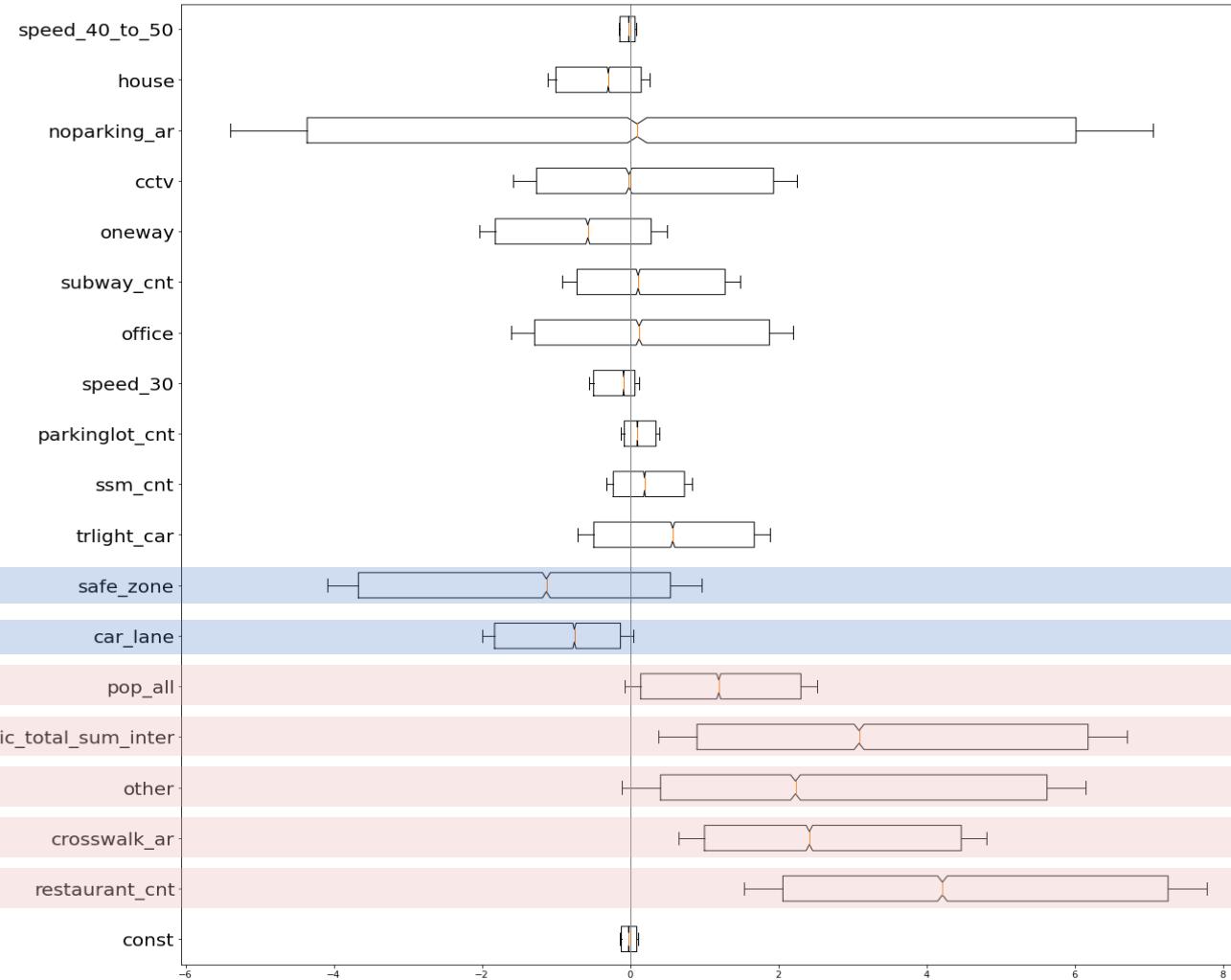
- 일방통행(oneway)
- 자동차 전용 도로(car\_lane)



교통량, 시설물, 횡단보도 면적 등의 변수가 양의 관계를 가지는 것으로 보아 인구 및 차량 유동성이 높은 지역이 사고 위험도가 높을 것으로 예상됨

또한 일방통행과 자동차 전용 도로가 사고 감소에 효과가 있을 것이라 판단됨

## 차대사람\_모형



## 영향을 많이 주는 요인

## 양의 관계를 가지는 요인

총 인구수(pop\_all)

교통량

(traffic\_total\_sum\_inter)

일반 건물(other)

\*주거, 사무용 건물 제외

횡단보도 면적(crosswalk\_ar)

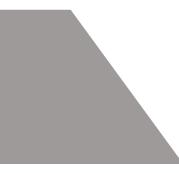
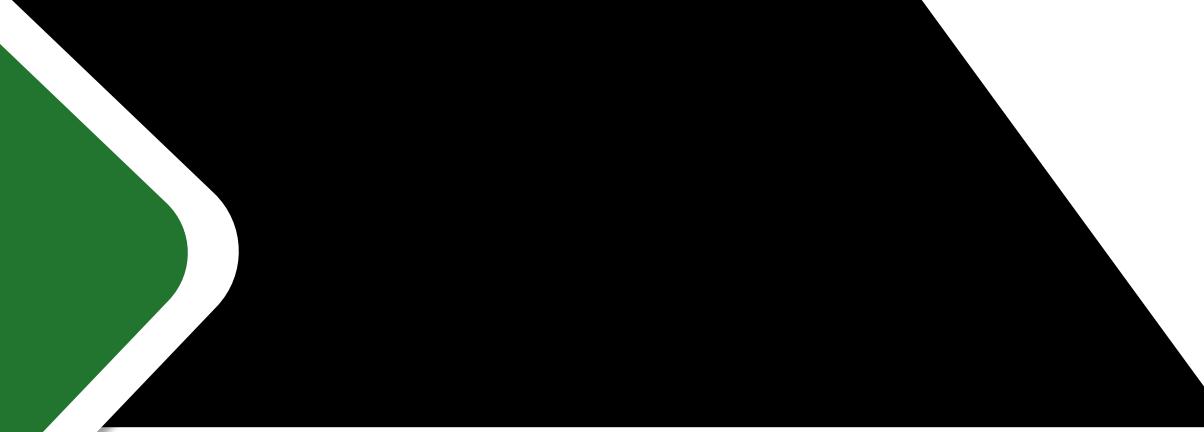
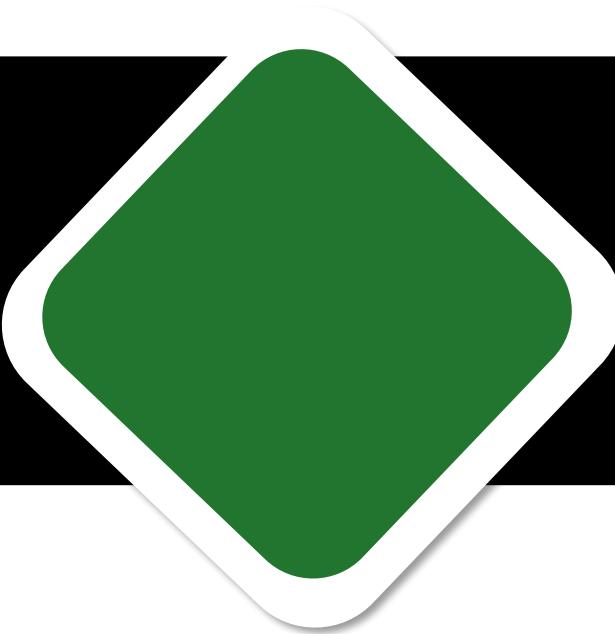
음식점(restaurant\_cnt)

## 음의 관계를 가지는 요인

안전지대(safe\_zone)  
자동차 전용 도로(car\_lane)

차대차 사고 모형과 유사하나 거주하는 인구가 높을수록 사고 증가에 영향을 끼치는 것을 알 수 있음

또한 안전지대가 사고 감소에 영향을 끼치는 것도 확인할 수 있음



# 위험지역 도출

## 위험지역 산정 기준

차대차 사고 모형과 차대사람 사고 모형의 예측값을 활용하여 위험지역 선정에 필요한 점수를 산출함

$$\text{Priority Score} = Y_i^{total} - (\widehat{Y_i^{car}} + \widehat{Y_i^{hum}})$$

$Y_i^{total}$  : i번째 격자의 전체 사고 위험도  
 $\widehat{Y_i^{car}}$  : i번째 격자의 차대차 사고 예측(by GWR 모형)위험도  
 $\widehat{Y_i^{hum}}$  : i번째 격자의 차대사람 사고 예측(by GWR 모형)위험도



Priority Score가 높을수록

- ( 1. 교통사고 개선이 시급한 지역 )
- ( 2. 정책적 교통사고 개선 효과가 우월한 지역 )

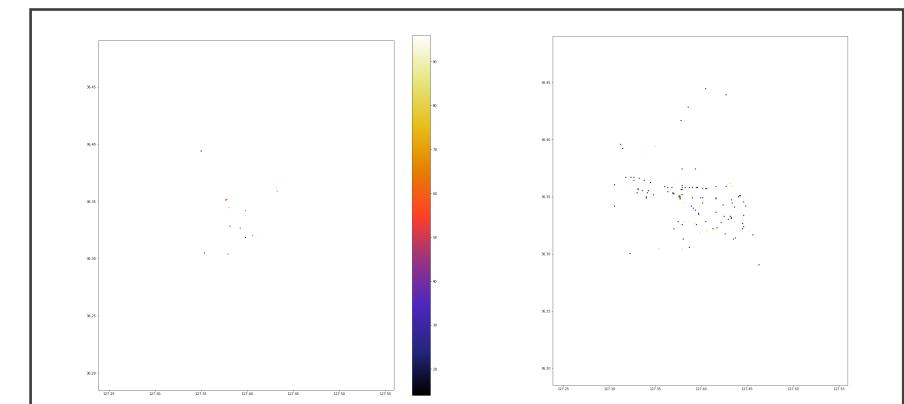


Priority Score가 높은 순으로 후보지 정렬 후 인접한 후보지를 하나로 묶어서 추천



K-means clustering 활용

: 후보지의 x, y 좌표 값으로 위치 정보 기반 클러스터링(n=100) 진행

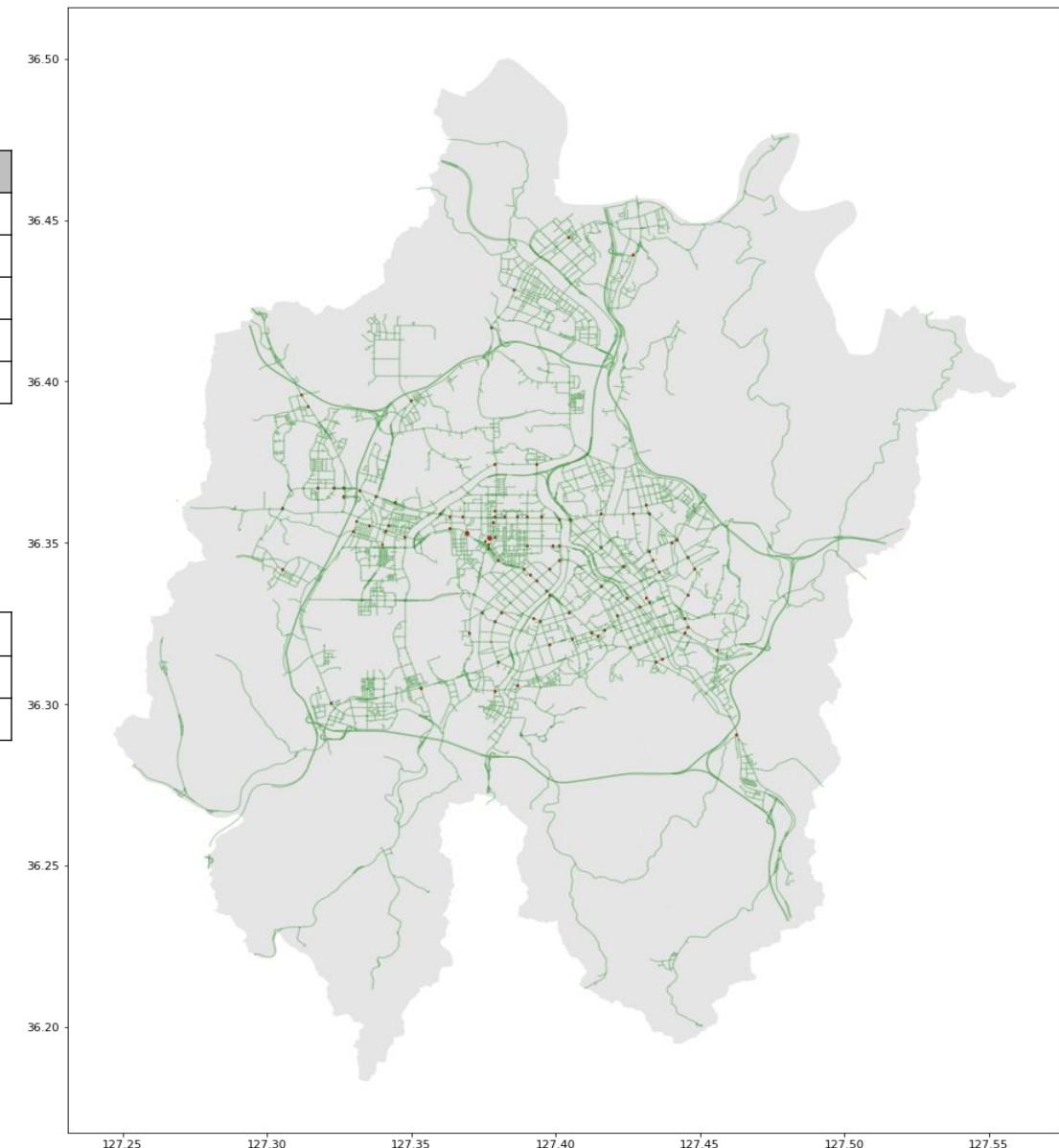


## 최종 위험지역 100개소

위험순위	시설명/주소지	X좌표(경도)	Y좌표(위도)	반경범위
1	덕명네거리	127.3054964	36.36061793	50
2	부사오거리	127.4348336	36.31297667	50
3	외삼네거리	127.3120996	36.3957884	50
4	계룡로 대전일보사 앞	127.3634682	36.35438746	50
5	중리네거리	127.4325661	36.35895414	50

•  
•  
•

98	배재로 축산농협 일대	127.3702093	36.32193944	50
99	유천네거리 스타벅스 일대	127.398066	36.31836025	50
100	변동네거리	127.3791164	36.32555496	50



### 최종 위험지역 100개소 - 예시



#### 위험순위 1) '덕명네거리'

- > 교차로, 횡단보도 등 사고위험도를 높이는 요인들이 눈에 띤다.
- > 주변에 아파트가 많은 것으로 보아 거주인구가 많은 지역으로 예상됨
- > 안전지대가 포함된 지역



#### 위험순위 2) '부사오거리'

- > 교차로, 횡단보도가 다수 있는 복잡한 도로
- > 주변 건물개수가 많음
- > 유동성이 큰 것으로 보임

결론

 요약차 대 차  
차대사람

교통량이 많거나 횡단보도가 많은 지역은 사고 위험에 크게 노출됨  
지하철역, 음식점 등 사회 인프라 시설이 밀접한 지역 또한 높은 위험도를 보임

반면, 자동차 전용도로에서는 낮은 위험도를 가지고 있음

## 차 대 차

CCTV가 많이 설치된 지역은  
차대차 사고의 위험도가  
높으나 인과관계가 있다고  
보기는 어려움

일방통행 도로에서는 차대차  
사고의 위험도가 낮음

## 차대사람

거주인구가 많은 지역은  
차대사람 사고 위험에 크게  
노출됨

안전지대를 포함한 지역에서  
낮은 위험도를 가지고 있음

## 차량단독

적은 사고건수

기존 변수로 설명이  
어려움

 시사점 도출

1.

## 교통사고 관련 다양한 정책적 근거로 사용 가능

지리적가중회귀분석에서는 각 **구역별**로 변수들의 영향이 다르게 적용되어 있고 안전시설물 외에도 도로제한속도나 차선, 자동차전용도로 등 **도로특성**에 따른 사고감소효과를 파악할 수 있기 때문에 **맞춤형 제도**를 실시하는 근거로 사용할 수 있음

같은 안전시설물을 설치하더라도 가장 효과가 높은 지역에 설치하는 등 같은 예산 안에서 사고감소효과를 극대화

2.

## 사고건수 대신 사고위험도를 산출해 분석에 사용했기 때문에 더 정밀한 비교가 가능



## 한계점

1.

대부분의 독립변수에 시계열 정보가 포함되어 있지 않아 변수들의 선형성 분석이 어려움

2.

컴퓨터의 가용 소스, 분석 시간 등 한정된 자원의 문제로 더 다양한 기법과 최적화를 적용하지 못함

3.

사용 가능한 변수가 더 많았다면 더 정밀한 분석이 가능했을 것임



## 추가분석방향

1.

'음주운전 사고여부', '시민들의 교통법규 준수 정도' 등의 다양한 변수 추가

2.

포아송 분포나 로지스틱 회귀모형 등의 다양한 기법을 지리적 가중회귀분석과 접목

# Reference



## 외부데이터

수집 데이터 셋	기준 연도	출처
parkinglot_cnt.csv		
restaurant_cnt.csv		
school_cnt.csv	2021	KAKAO DEVELOPER
ssm_cnt.csv		
subway_cnt.csv		



## 참고논문

- 공간의존성에 대한 이해와 공간회귀분석의 활용, 이석환(2014)
- 지리시간가중 회귀모형을 이용한 주택가격 영향요인 분석, 박세희(2017)
- 교통사고비용과 EPDO에 근거한 사고밀도 청주사례 분석, 박나영 , 박병호(2018)
- 도로교통사고 예방 위한 위험도 평가기법 모색, 홍상연, 정재훈(2018)
- mgwr: A Python implementation of multiscale geographically weighted regression for investigating process spatial heterogeneity and scale, Taylor M. Oshan (2018)

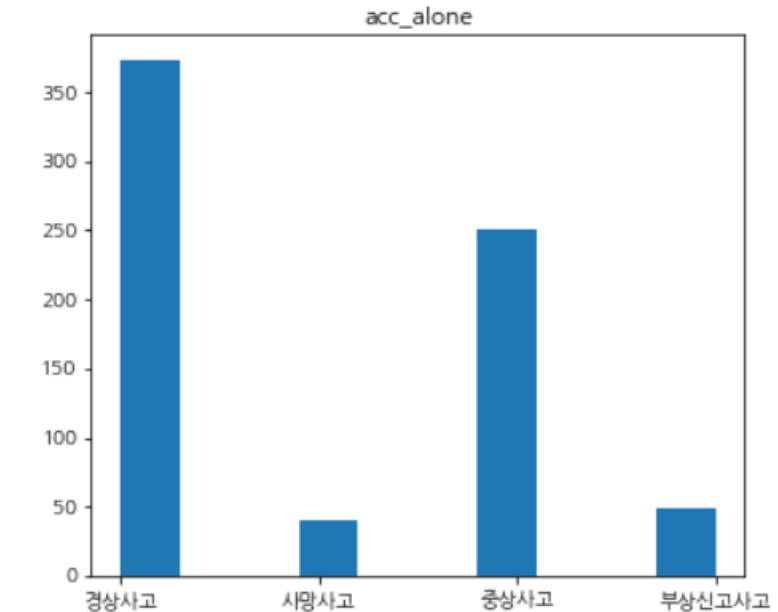
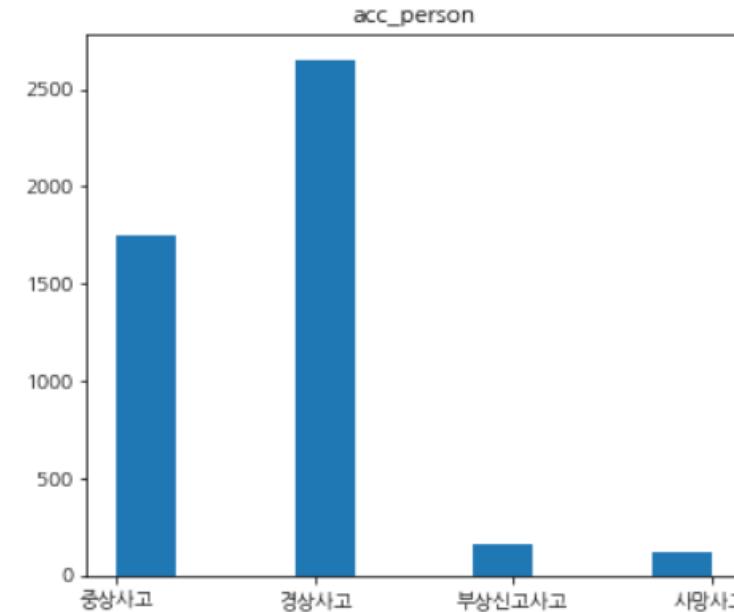
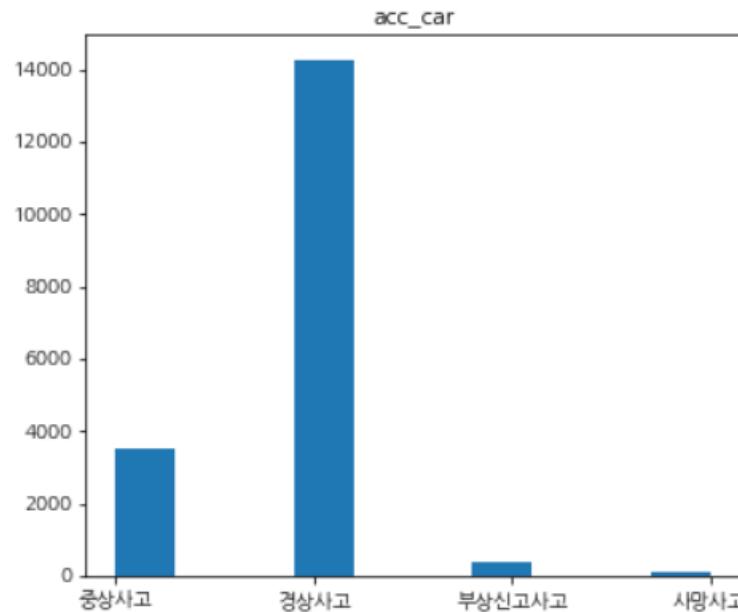
분석도구



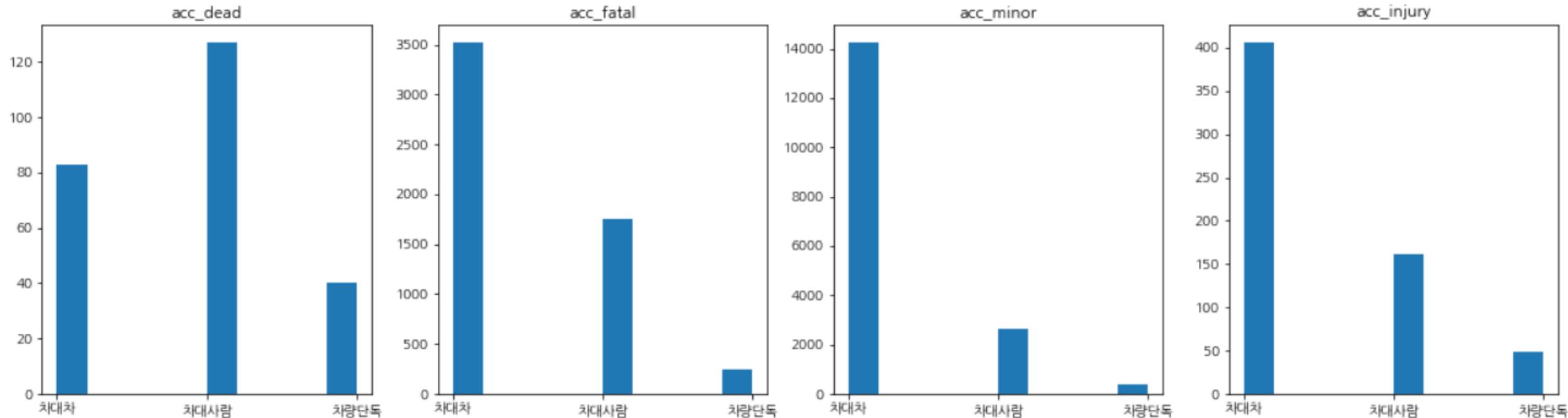
kakao developers

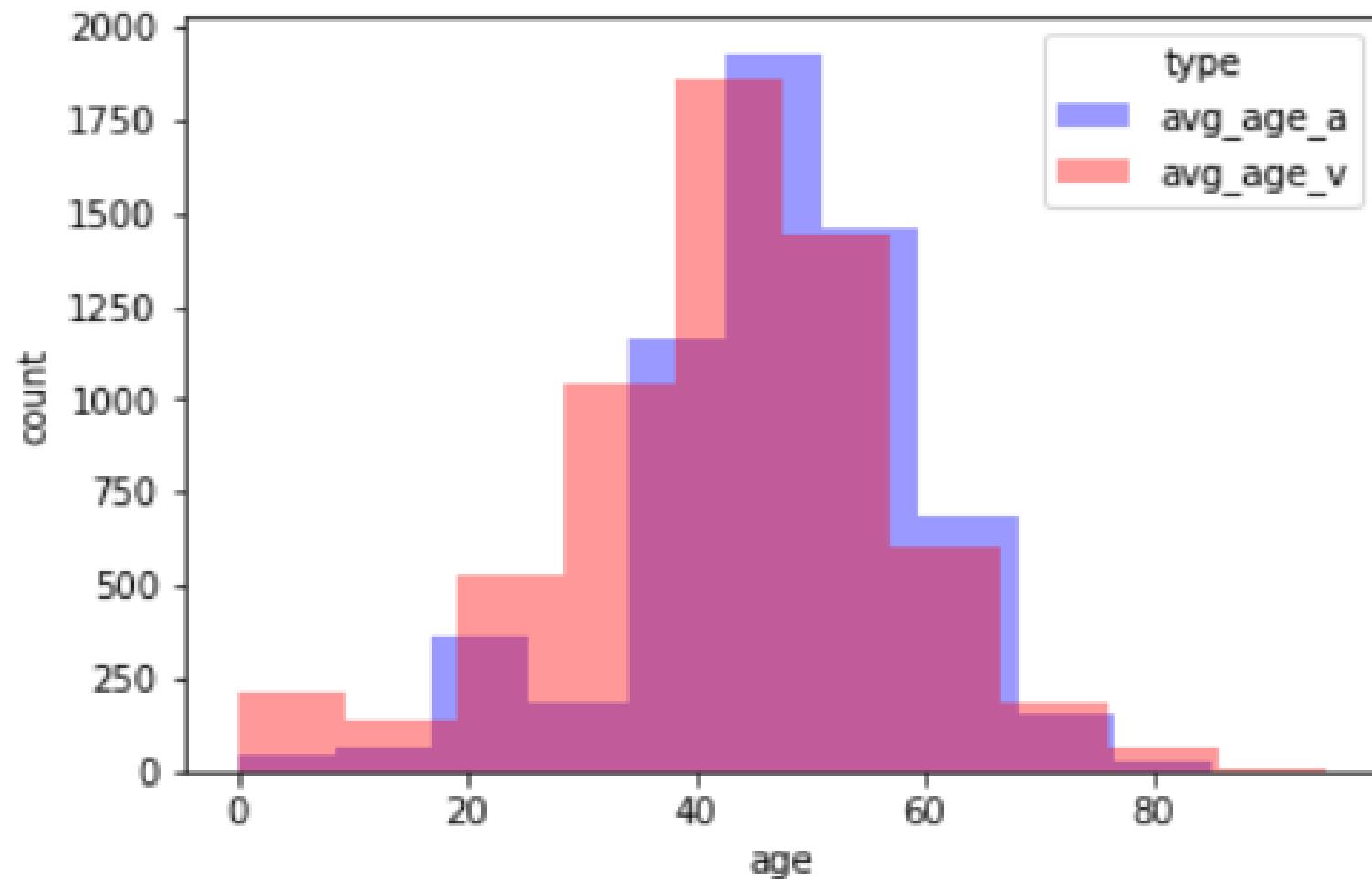
# 부록

[그래프1] 슬라이드9, 슬라이드25: 사고유형별 특징파악

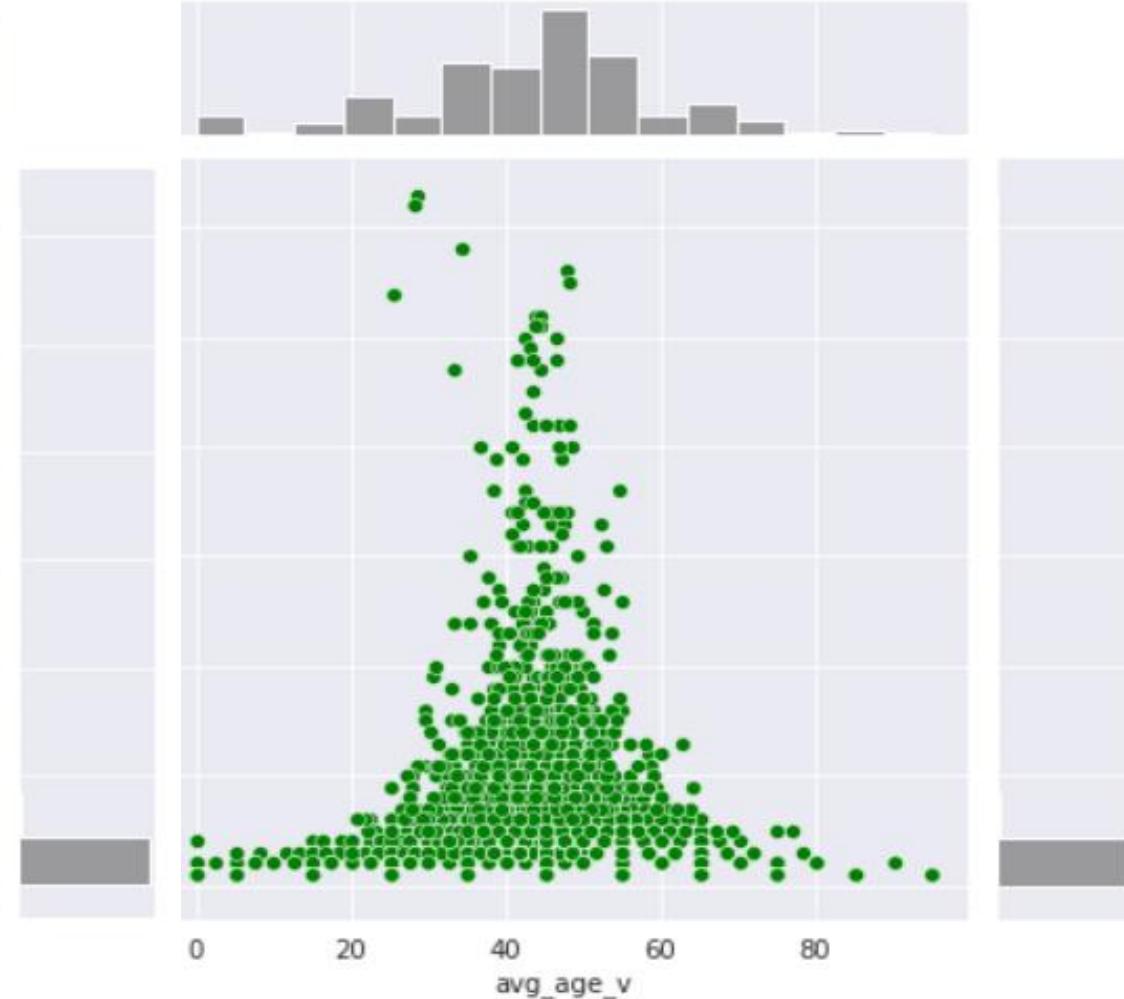
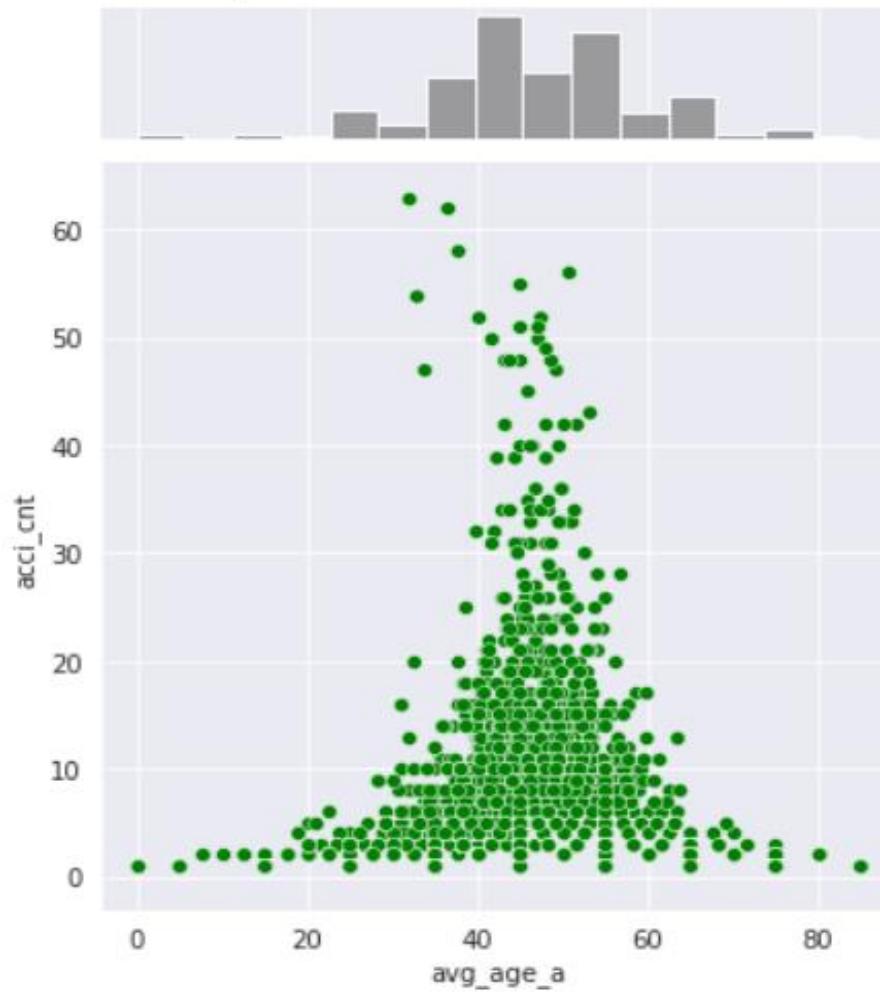


[그래프2] 슬라이드9: 사고유형별 특징파악

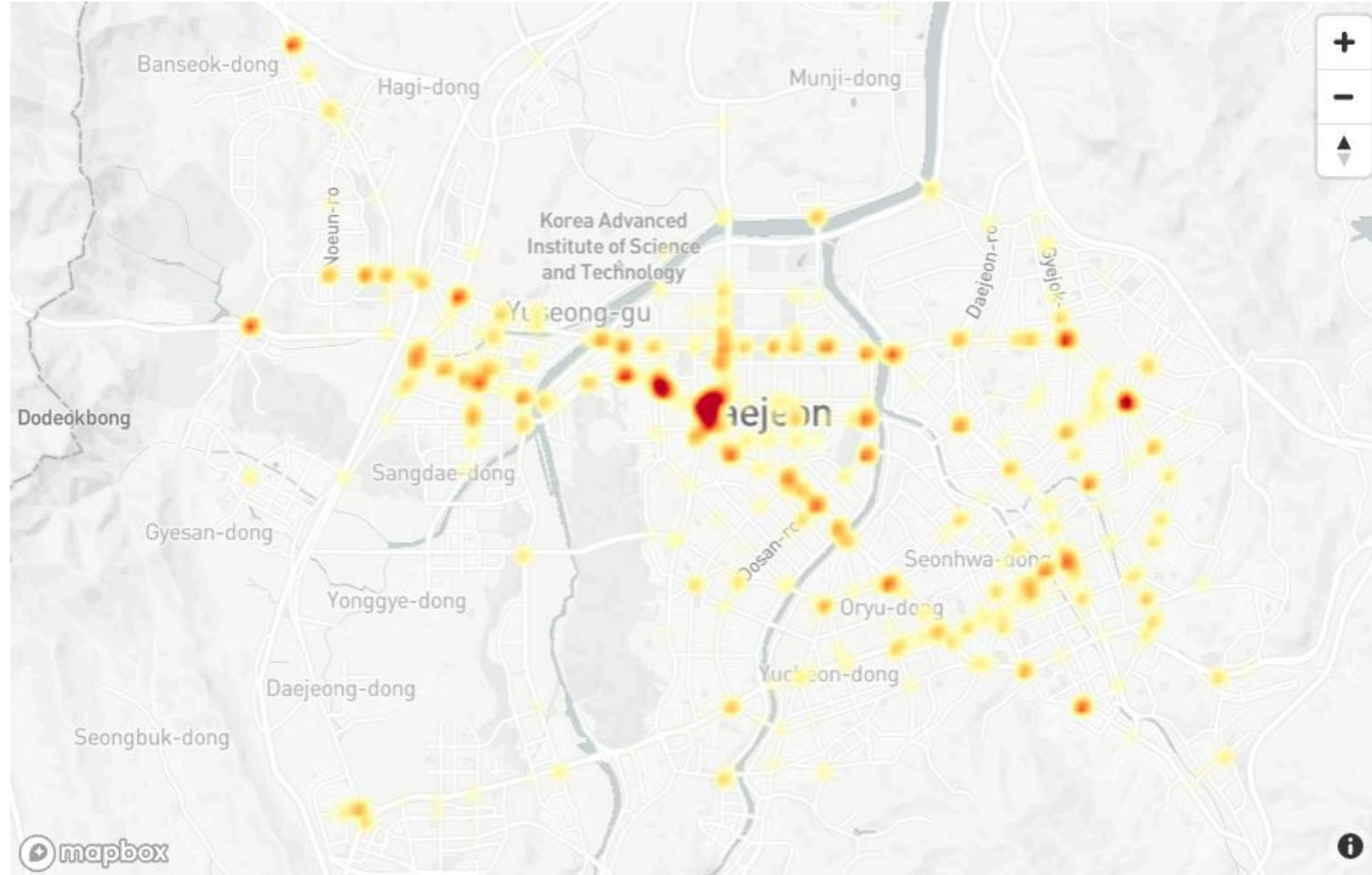


[그래프3] 슬라이드10: 연령대별 특징파악

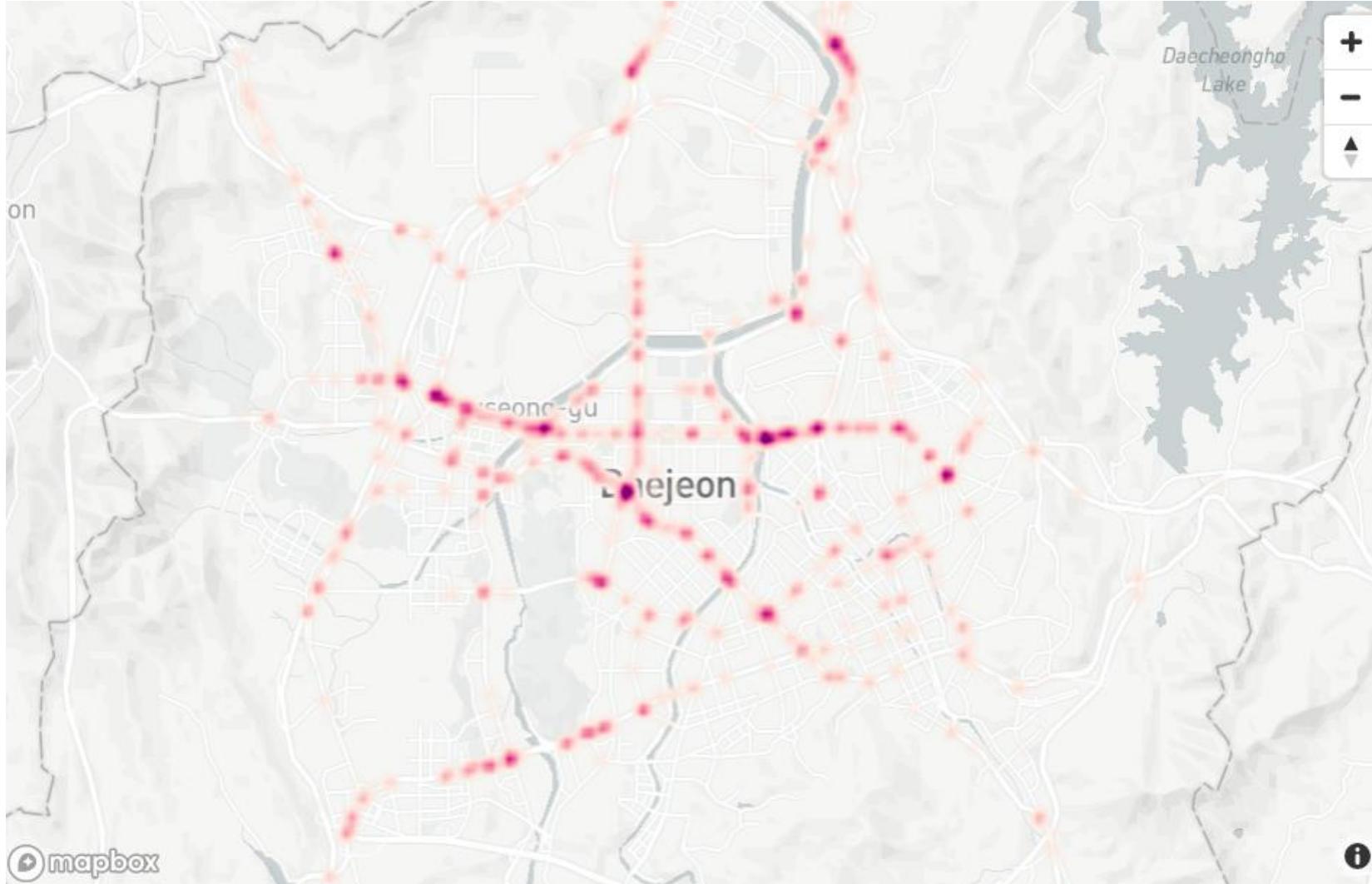
[그래프4] 슬라이드10: 연령대별 특징파악



[그래프5] 슬라이드11: 교통사고 데이터 시각화



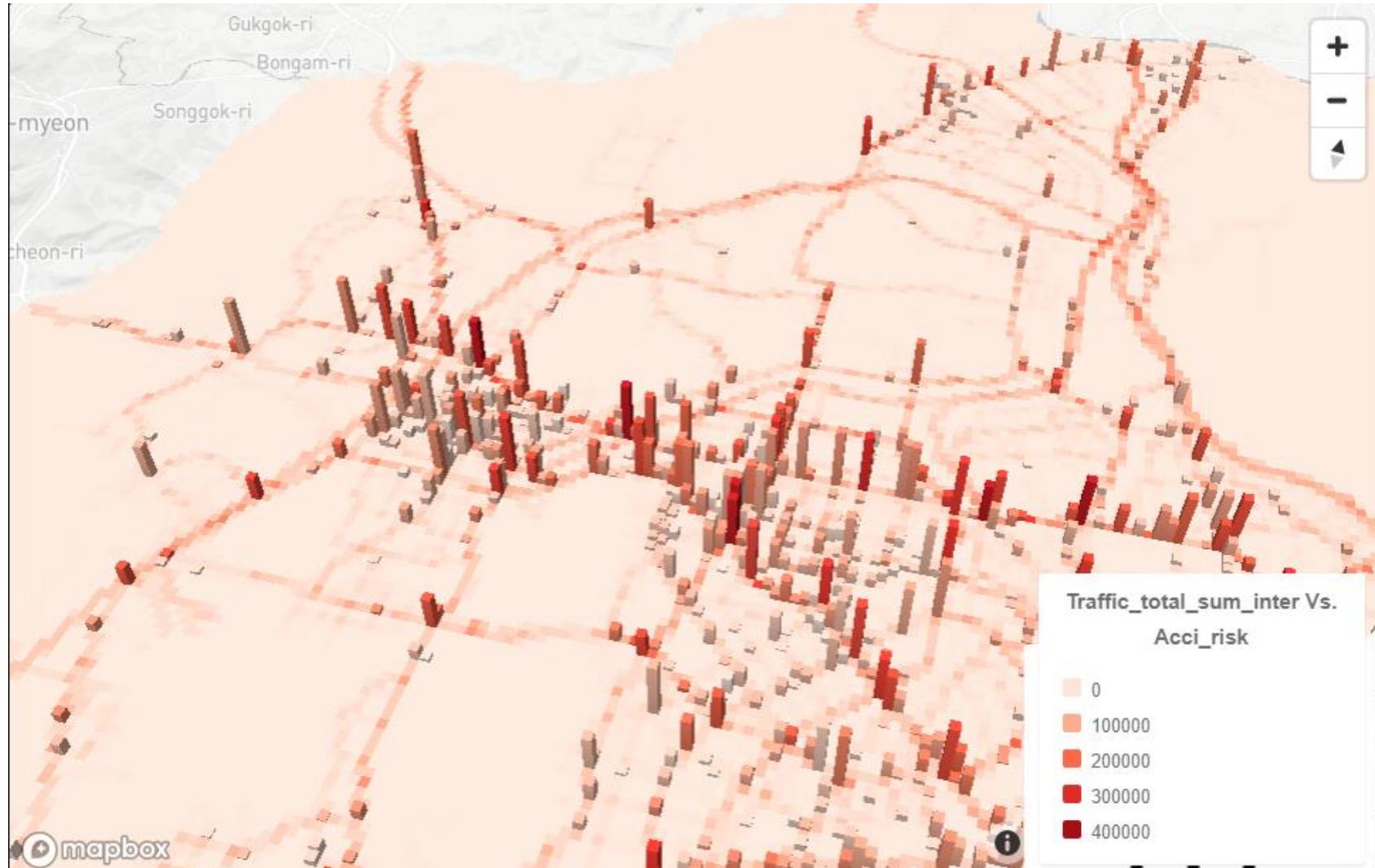
[그래프6] 슬라이드11: 교통량 데이터 시각화



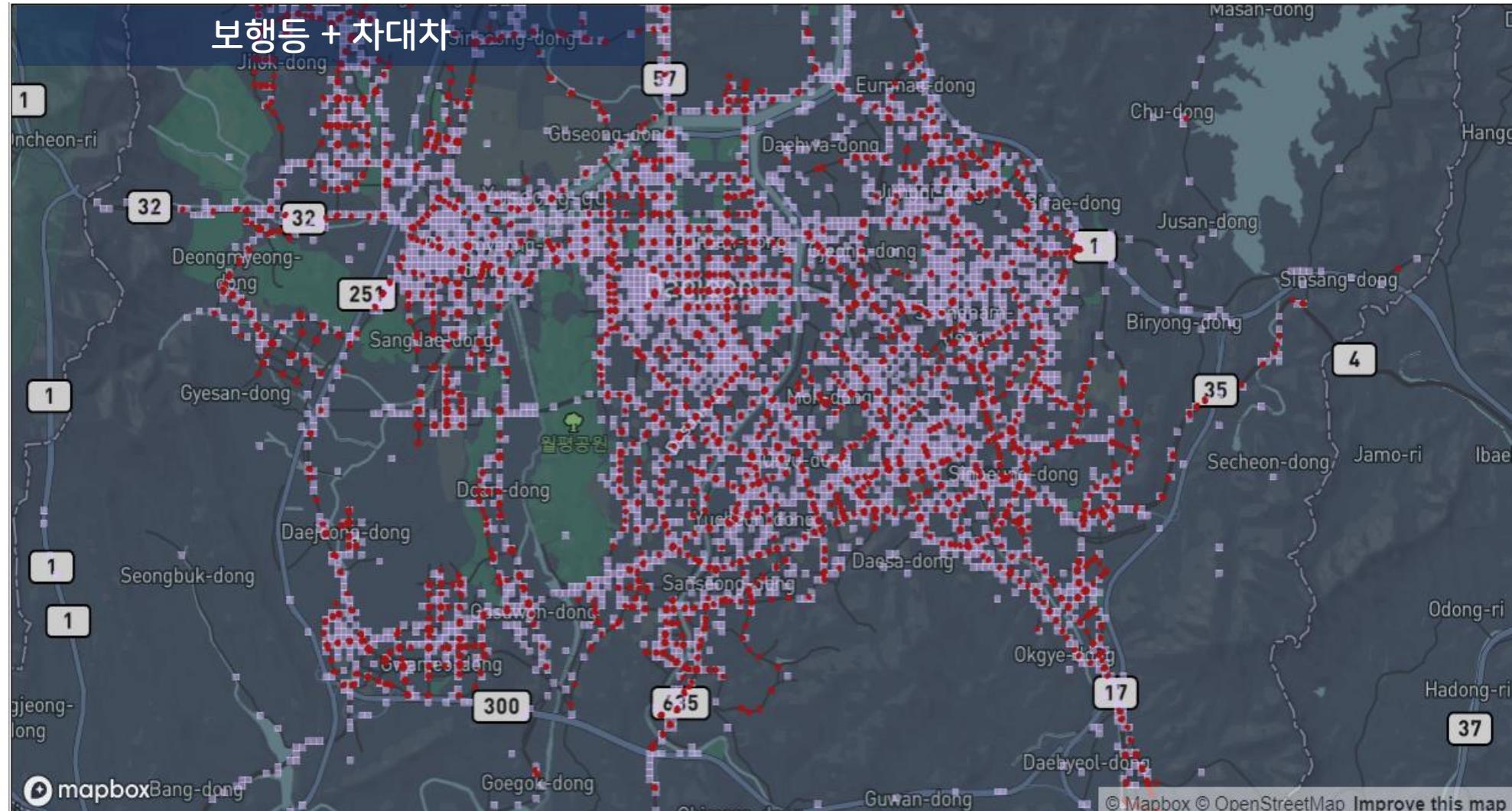
[그래프7] 슬라이드12: 사고건수 + 교통량 데이터 시각화



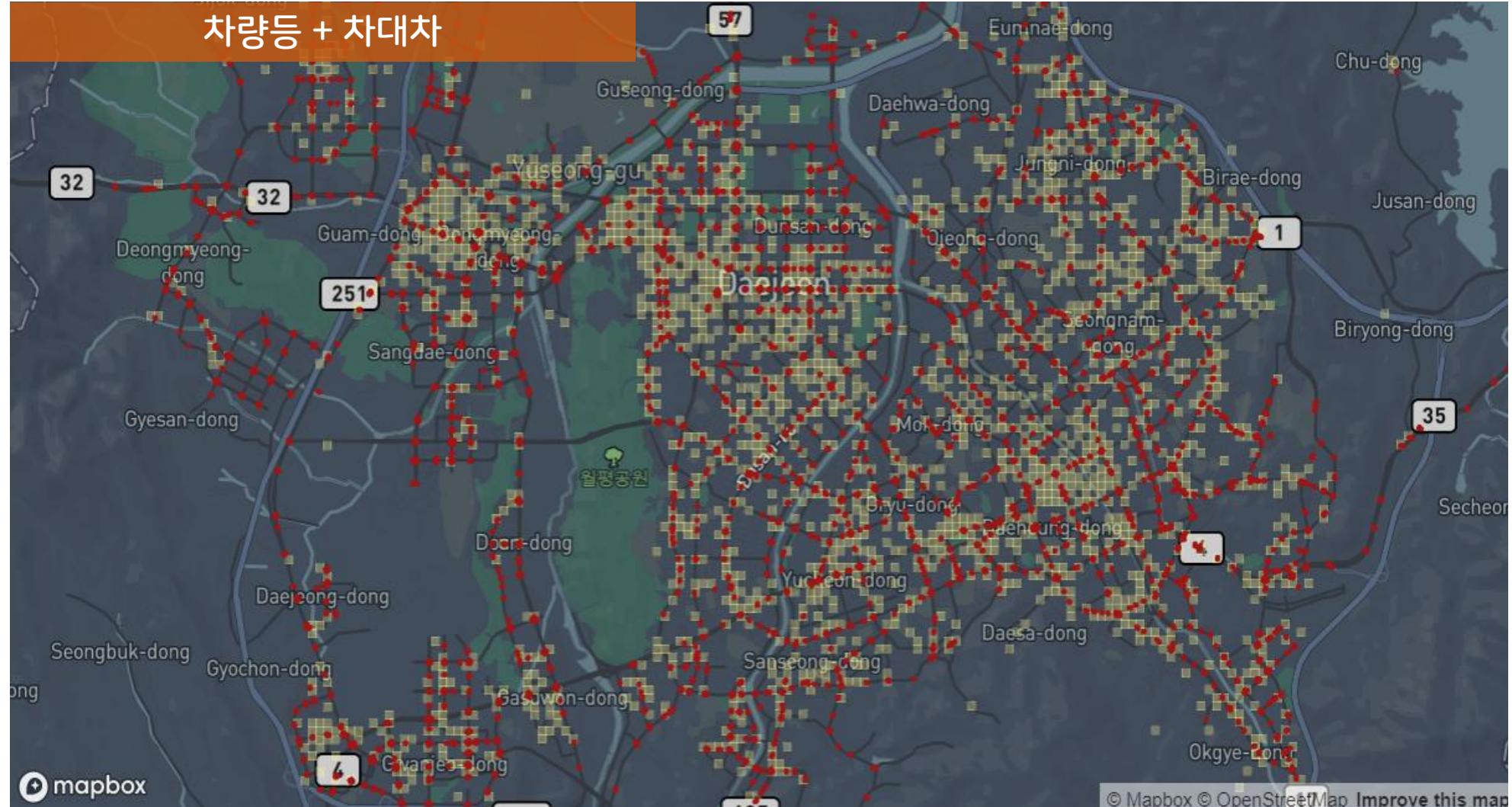
[그래프8] 슬라이드12: 위험도(EPDO지수) + 교통량 데이터 시각화



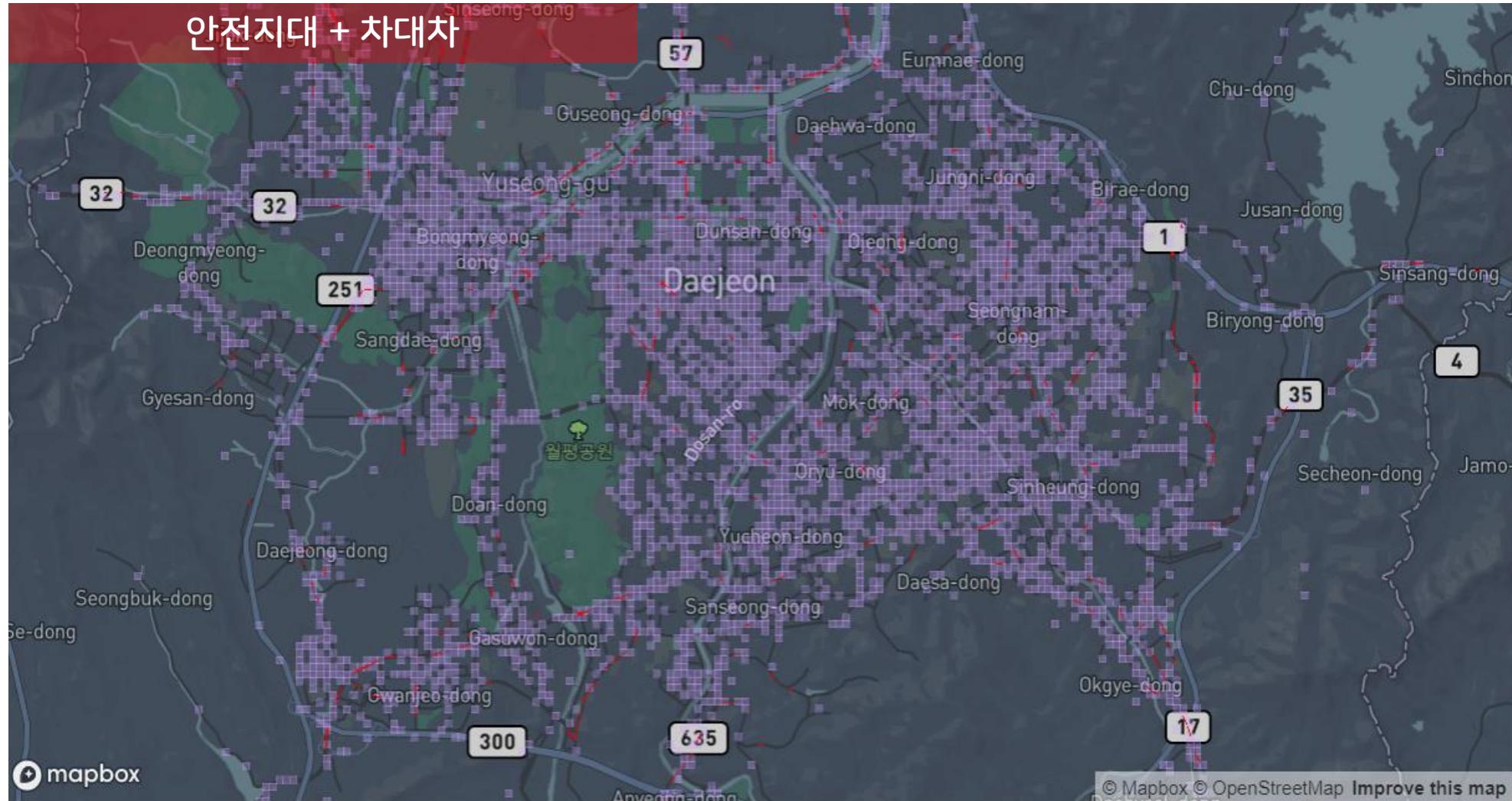
[그래프9] 슬라이드13: 보행등 + 차대차 사고 데이터 시각화



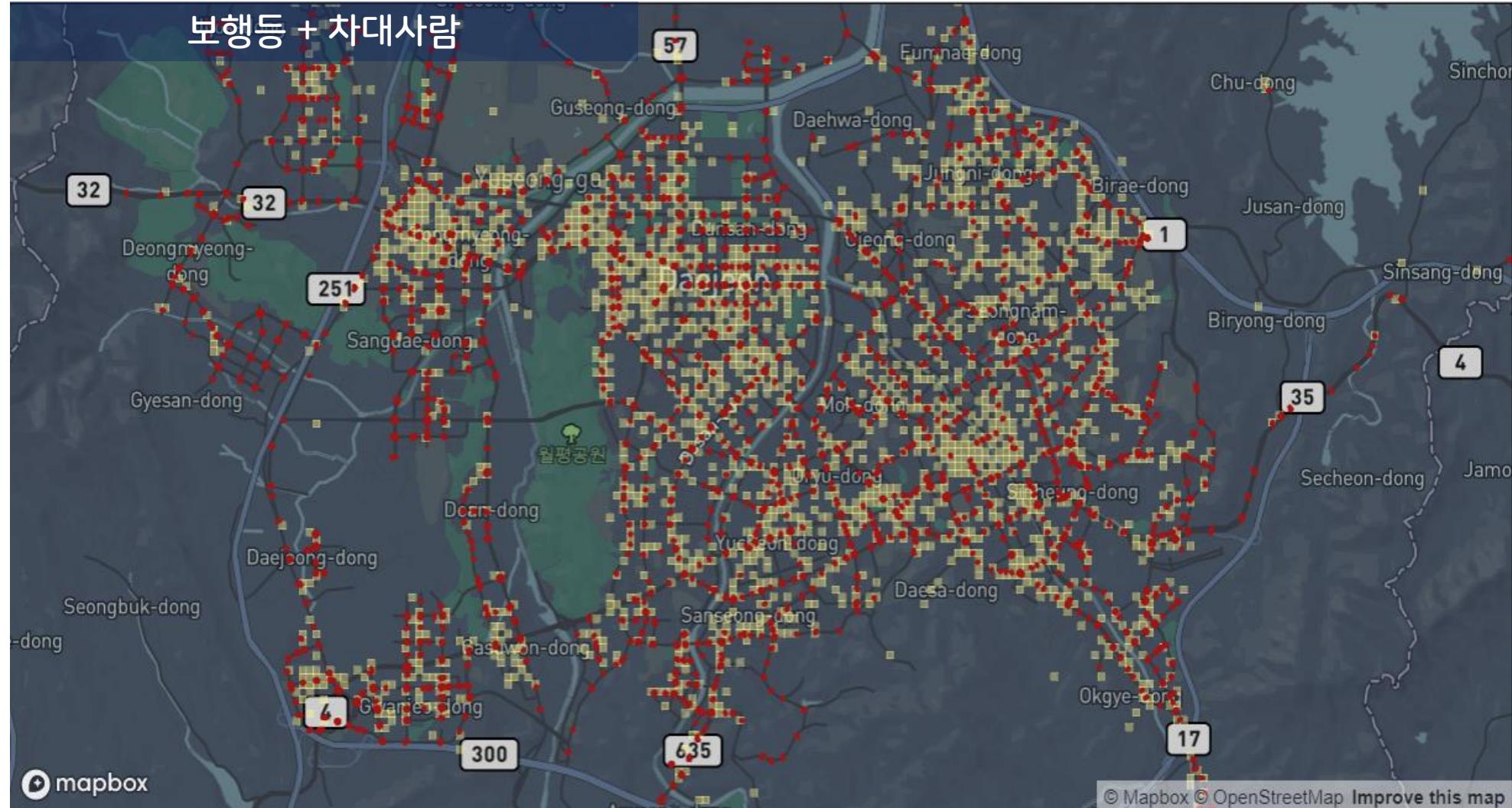
[그래프10] 슬라이드13: 차량등 + 차대차 사고 데이터 시각화



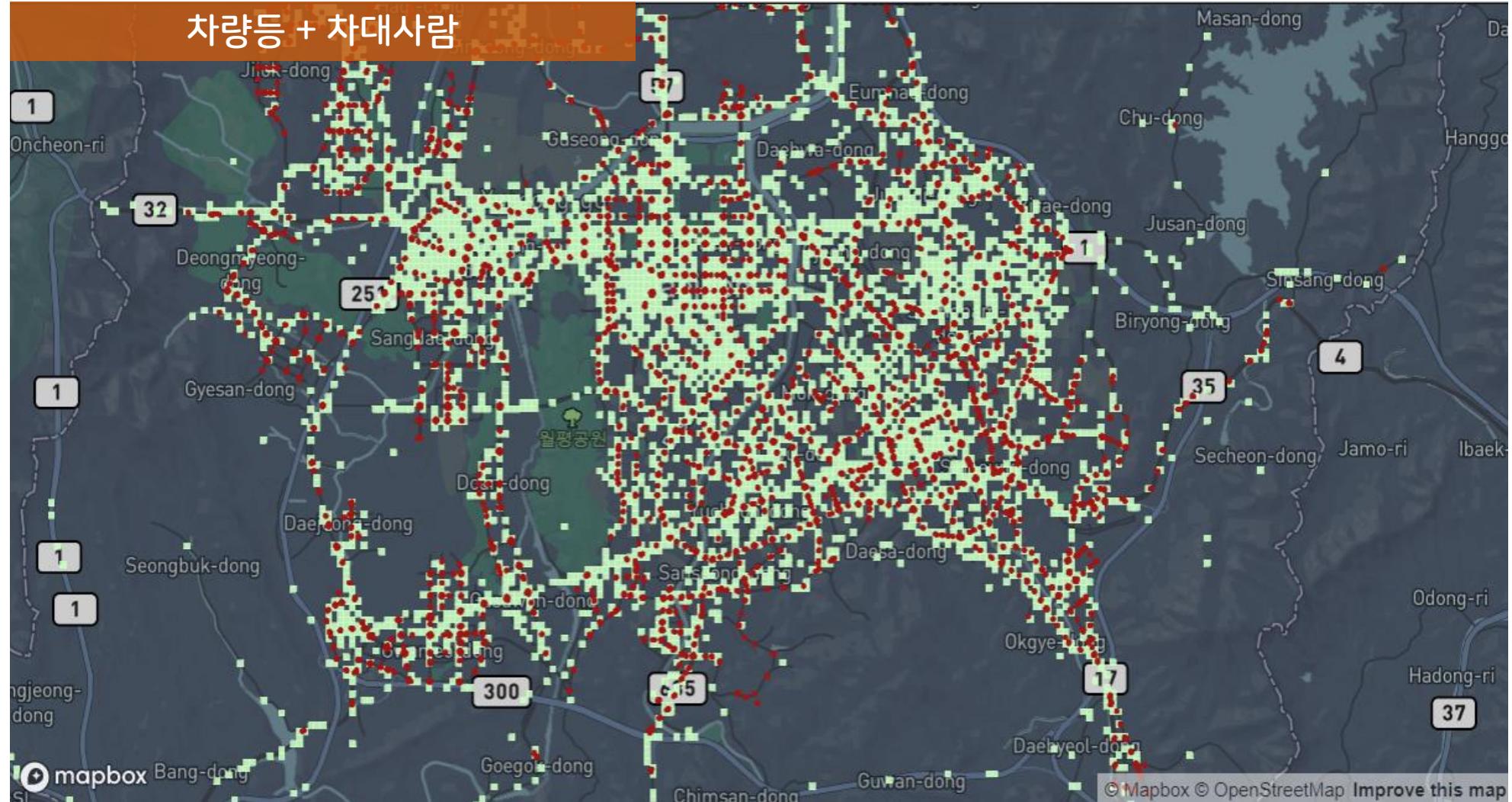
[그래프11] 슬라이드13: 안전지대 + 차대차 사고 데이터 시각화



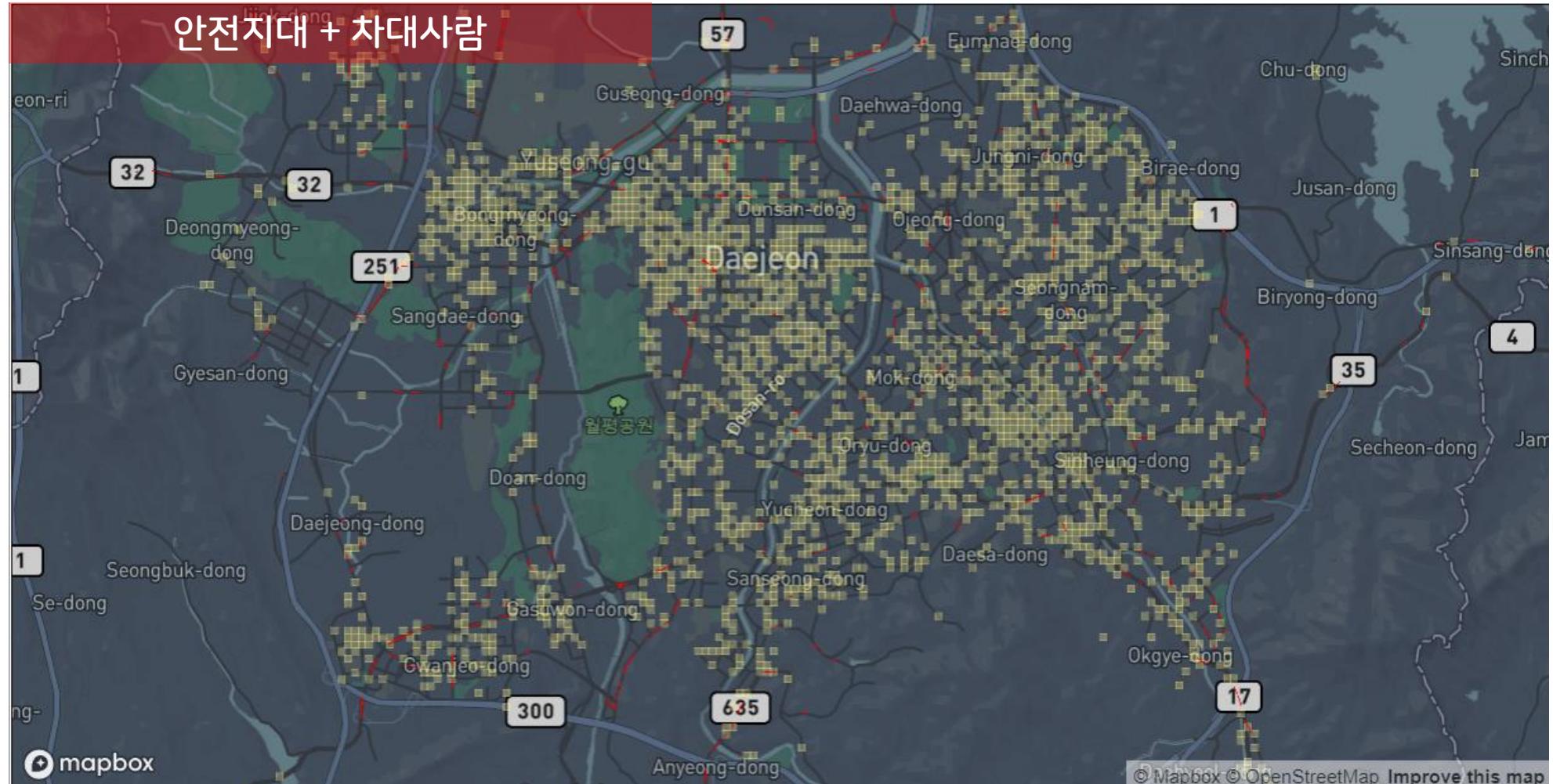
[그래프12] 슬라이드13: 보행등 + 차대사람 사고 데이터 시각화



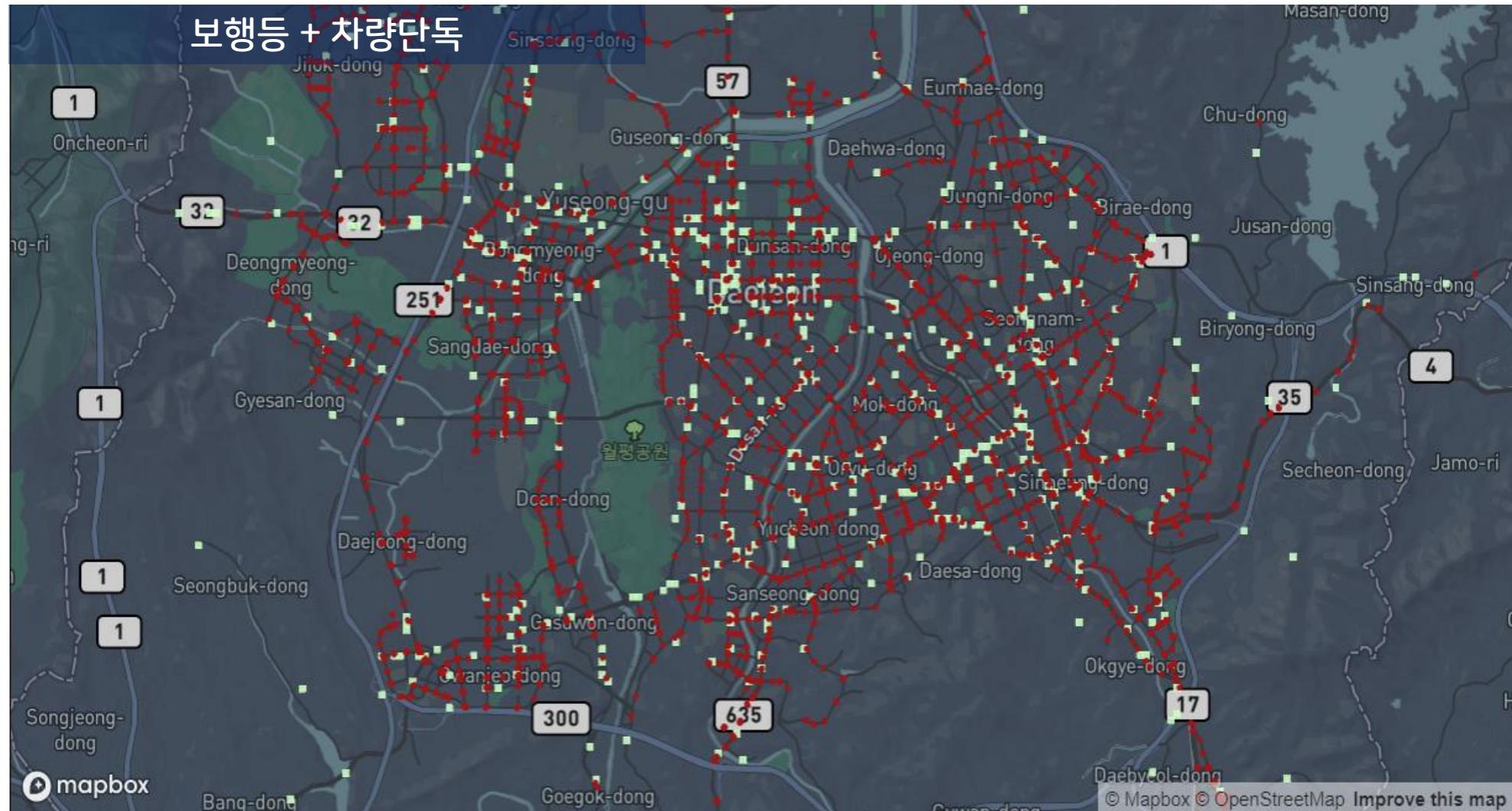
[그래프13] 슬라이드13: 차량등 + 차대사람 사고 데이터 시각화



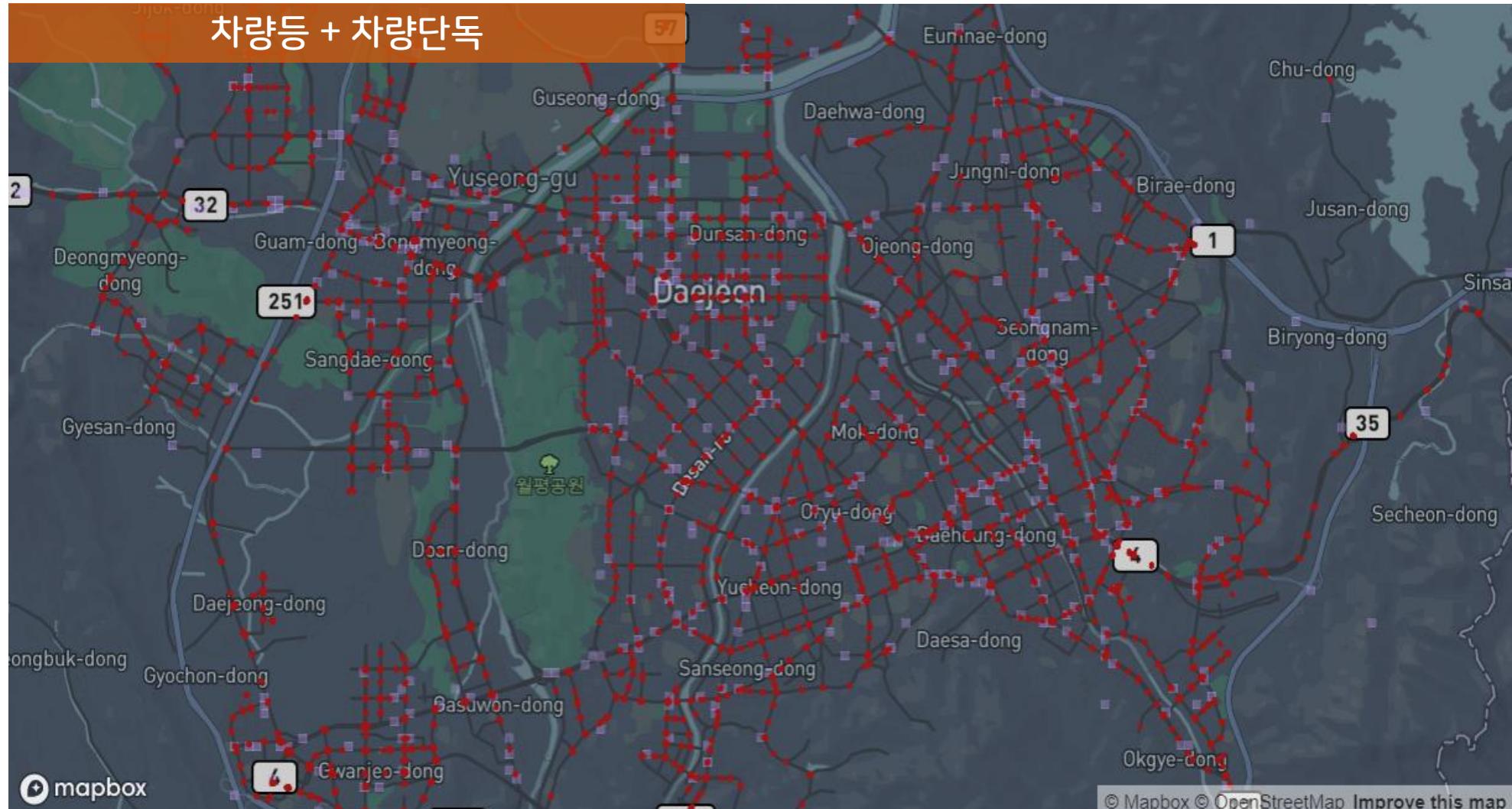
[그래프14] 슬라이드13: 안전지대 + 차대사람 사고 데이터 시각화

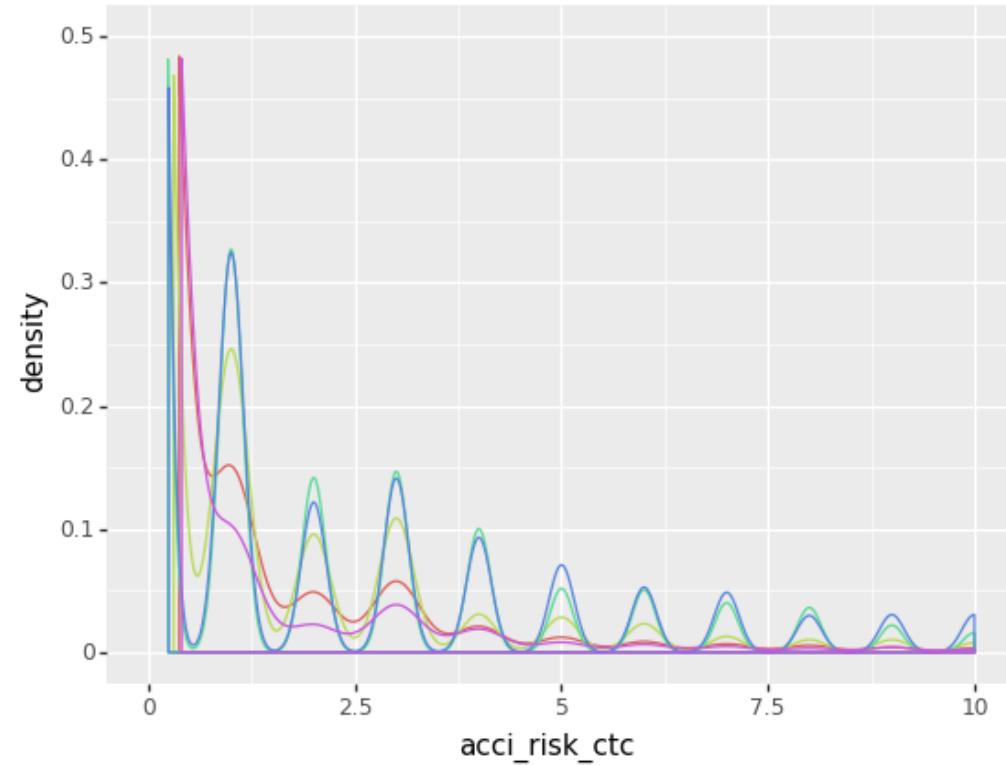


[그래프15] 슬라이드13: 보행등 + 차량단독 사고 데이터 시각화



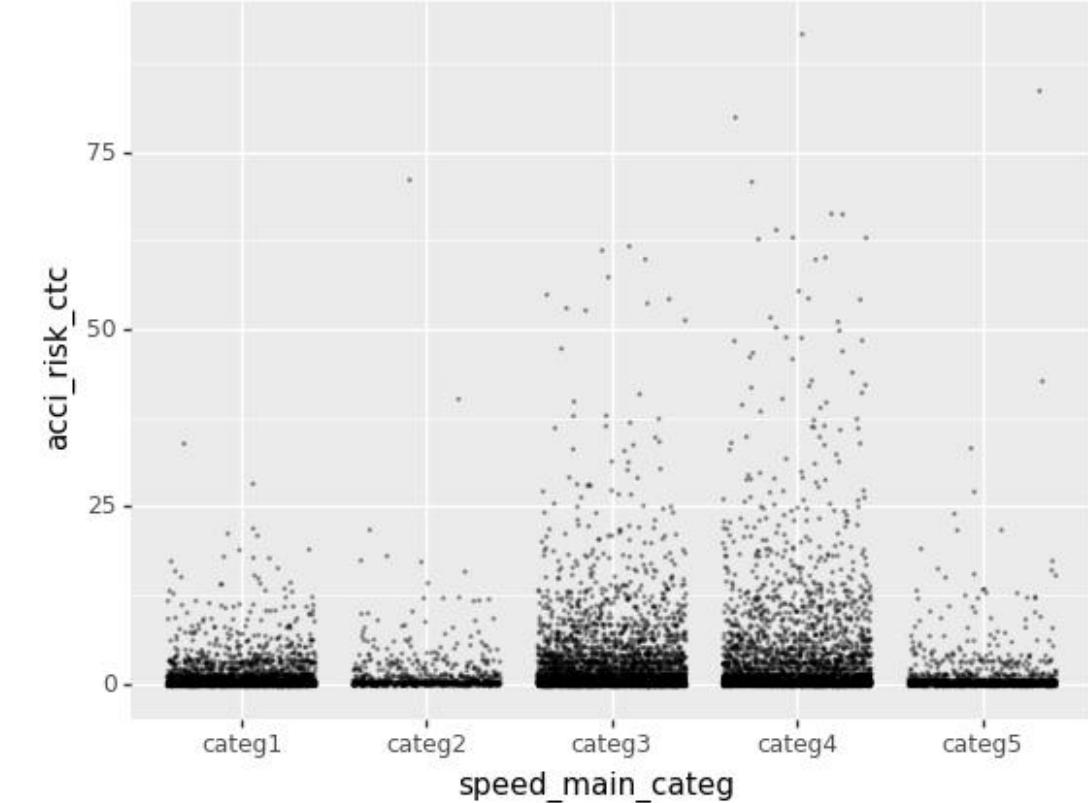
[그래프16] 슬라이드13: 차량등 + 차량단독 사고 데이터 시각화



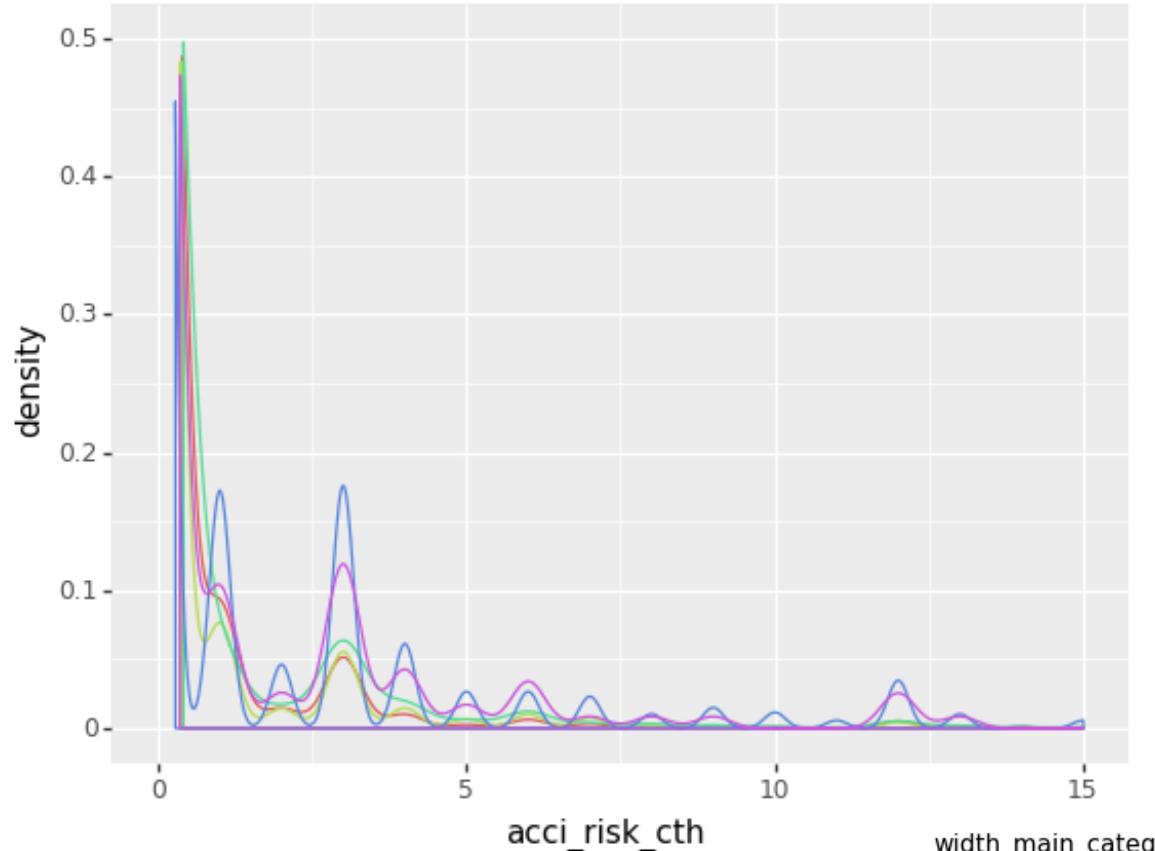
[그래프17] 슬라이드19: 속도제한변수 범주화 관련

speed\_main\_categ

- categ1
- categ2
- categ3
- categ4
- categ5



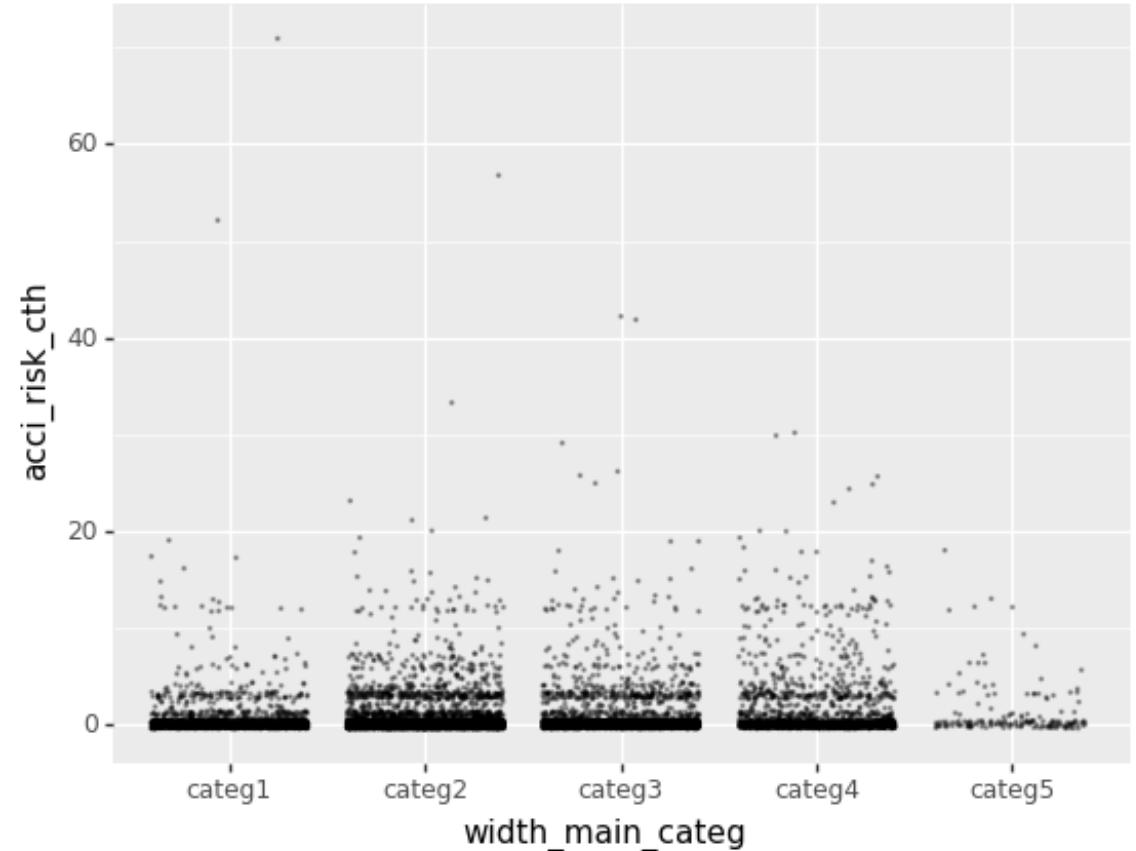
categ1: speed\_0 (속도 제한 표시가 없는 도로)  
categ2: speed\_30  
categ3: speed\_40\_to\_50  
categ4: speed\_60\_to\_70  
categ5: speed\_80\_to\_110

[그래프18] 슬라이드19: 차선변수 범주화 관련

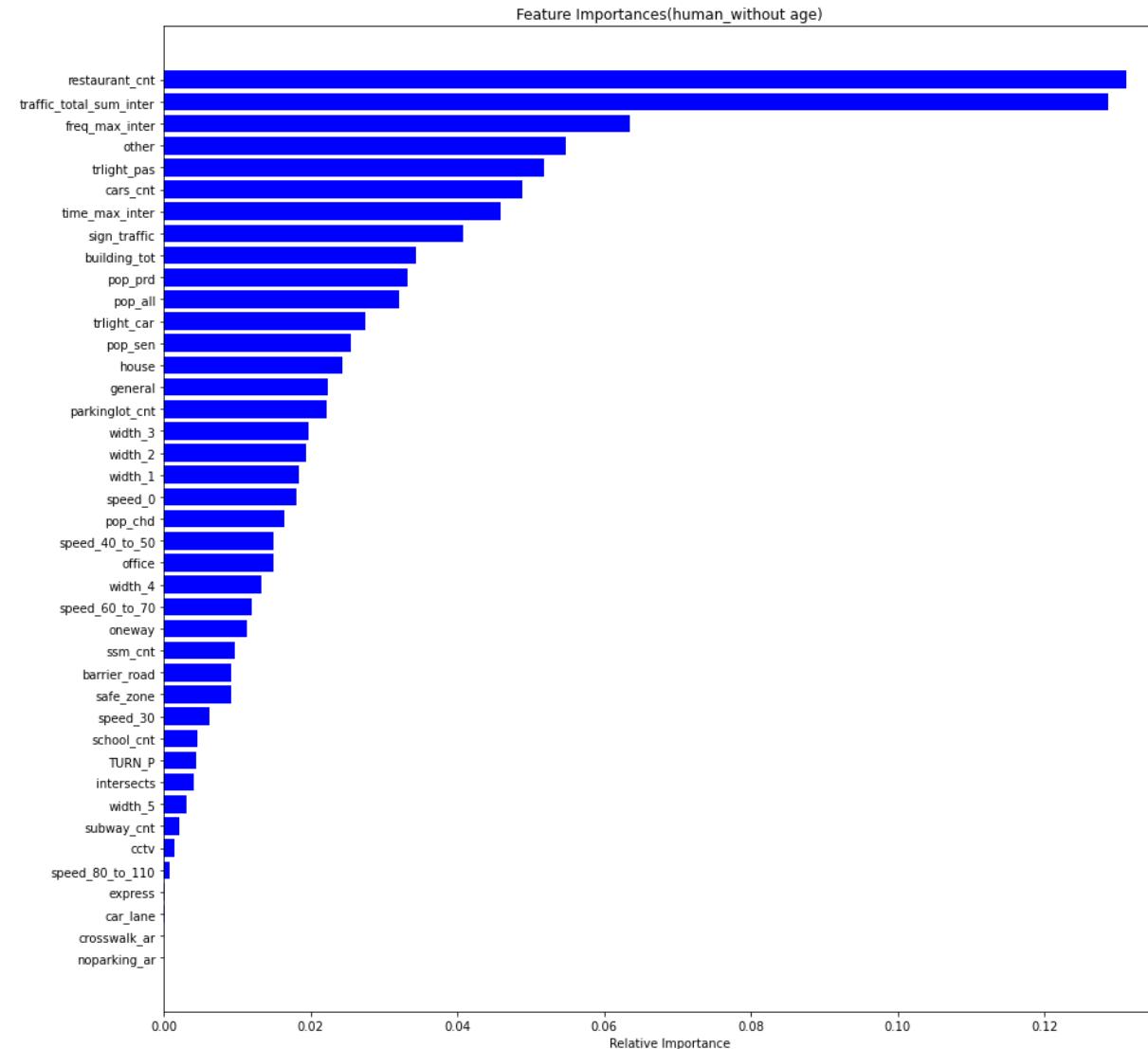
width\_main\_categ

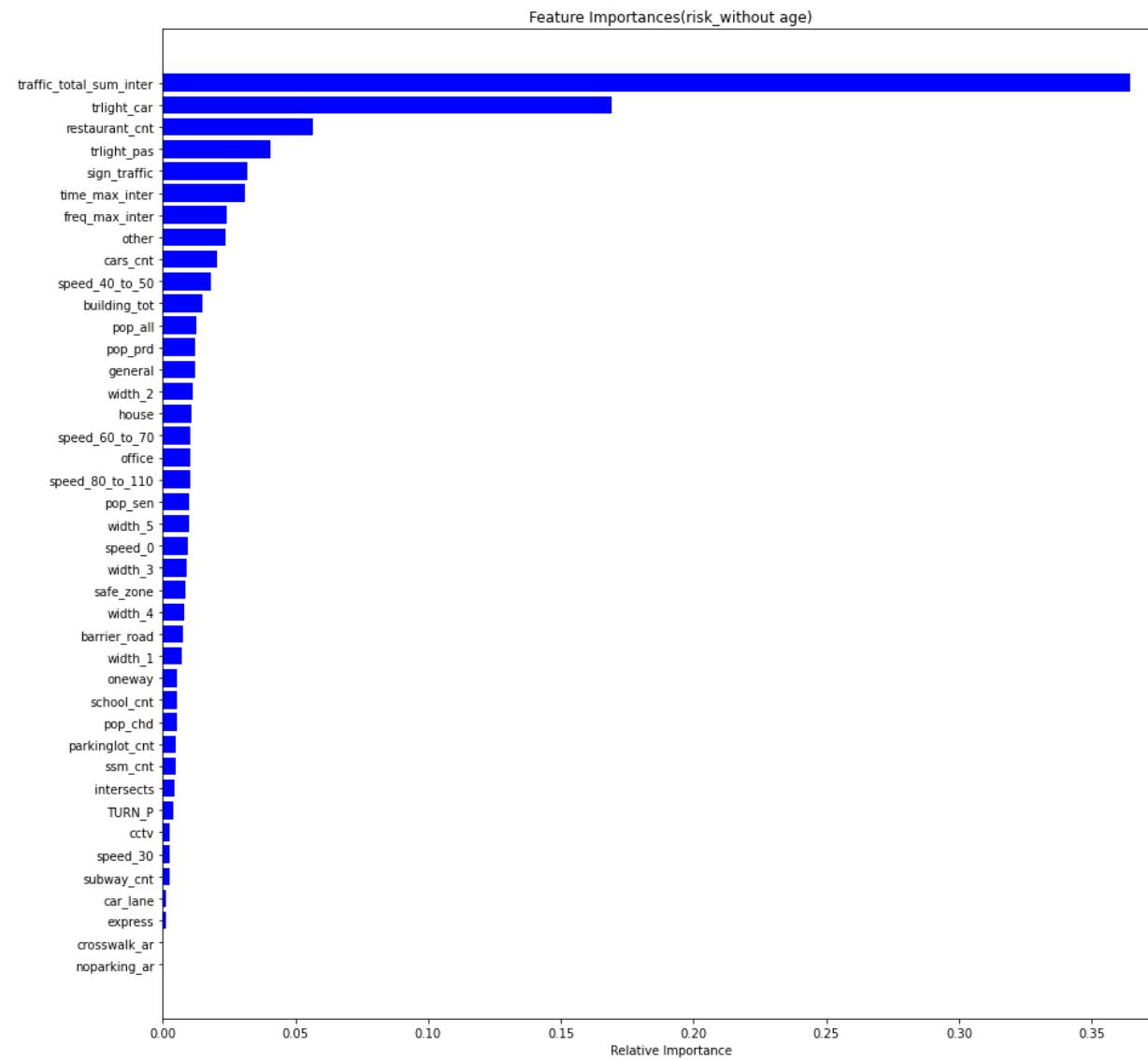
- categ1
- categ2
- categ3
- categ4
- categ5

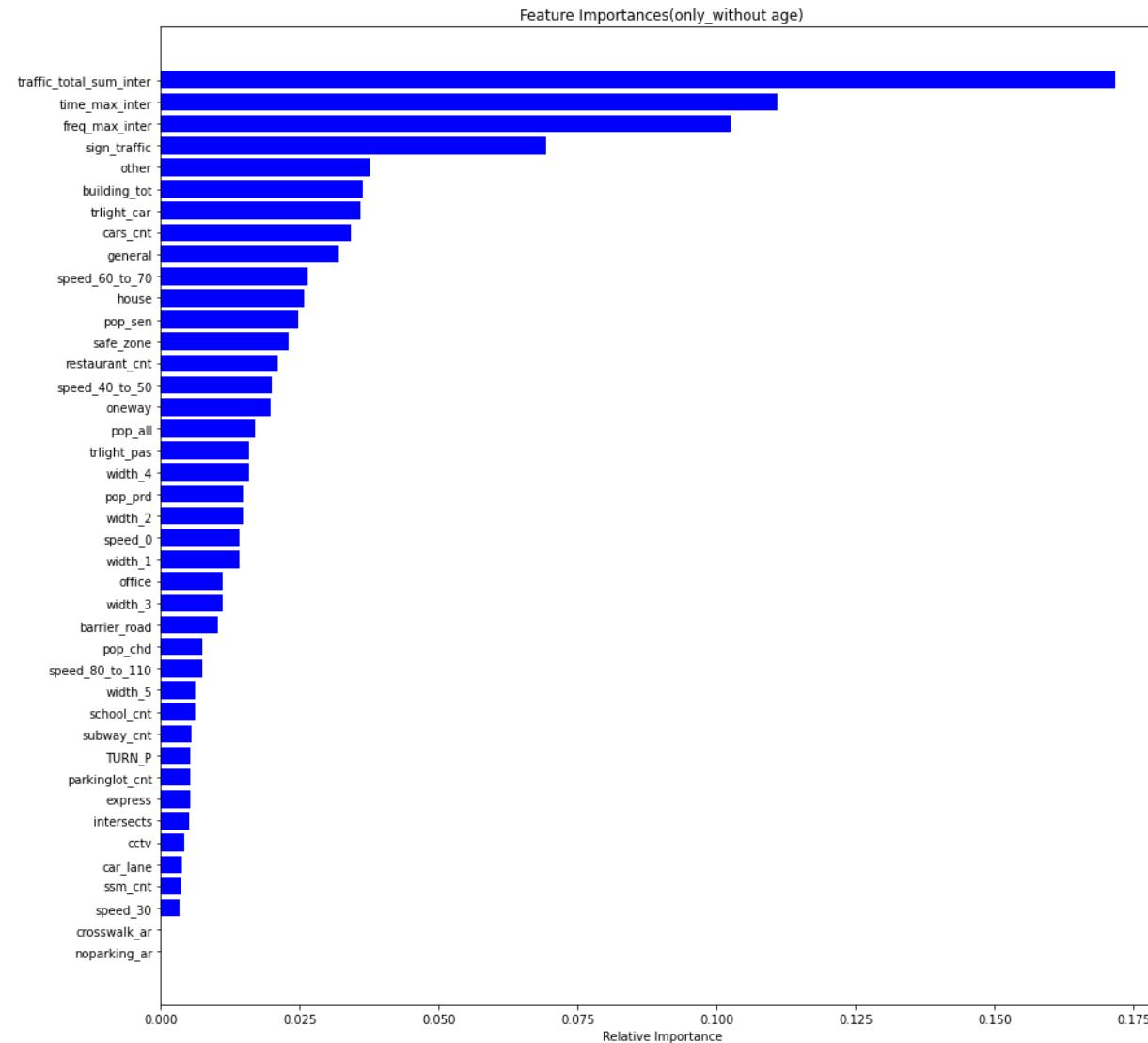
categ1: width\_1 (1차선)  
categ2: width\_2 (2차선)  
categ3: width\_3 (3차선)  
categ4: width\_4 (4,5,6,7,8차선)  
categ5: width\_5 (9차선 이상)



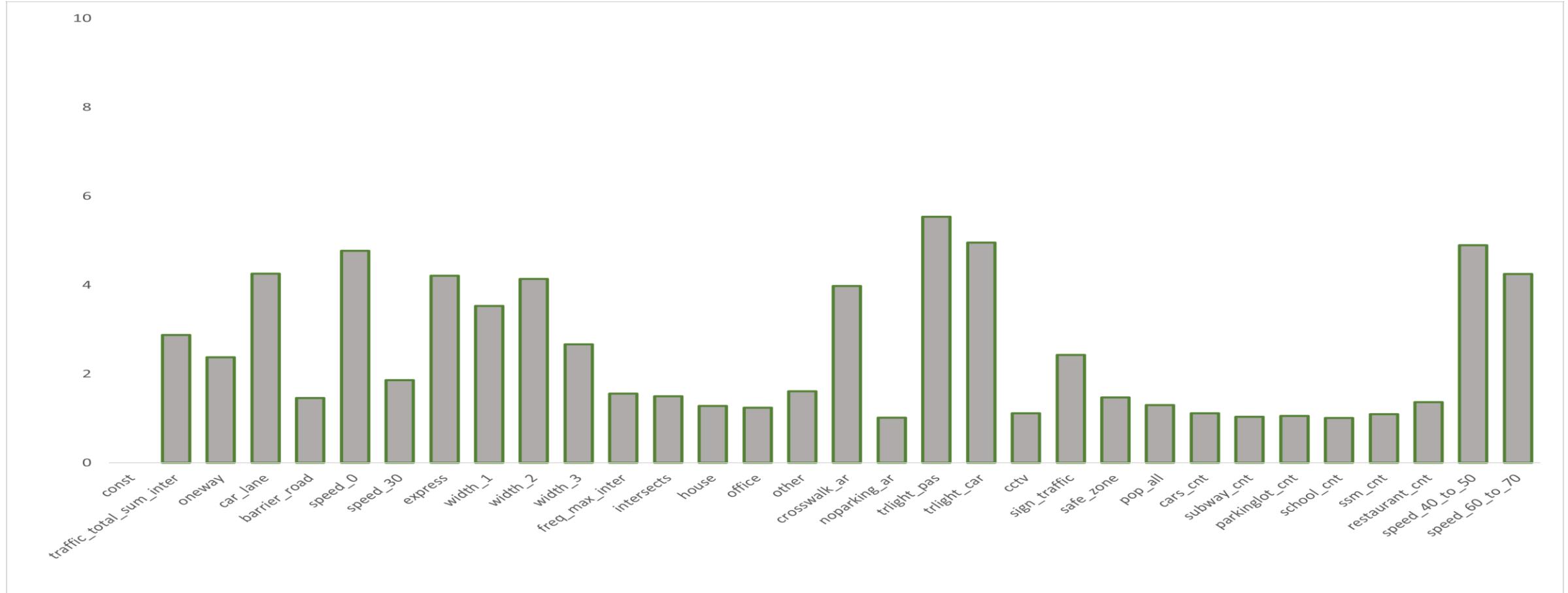
width\_main\_categ

[그래프19] 슬라이드25: RandomForest Feature Importance - 차대차 사고

[그래프20] 슬라이드25: RandomForest Feature Importance - 차대사람 사고

[그래프21] 슬라이드25: RandomForest Feature Importance - 차량단독 사고

[그래프22] 슬라이드31: 변수 제거 전 다중공선성

[그래프23] 슬라이드31: 변수 제거 후 다중공선성

[그래프24] 슬라이드45: k-means 클러스터링 결과

