

2021.04.01

Team 넷이서 5인분

대전시 교통사고 위험지역 도출

- 1 ————— 분석 개요
- 2 ————— 분석 과정
- 3 ————— 분석 결과

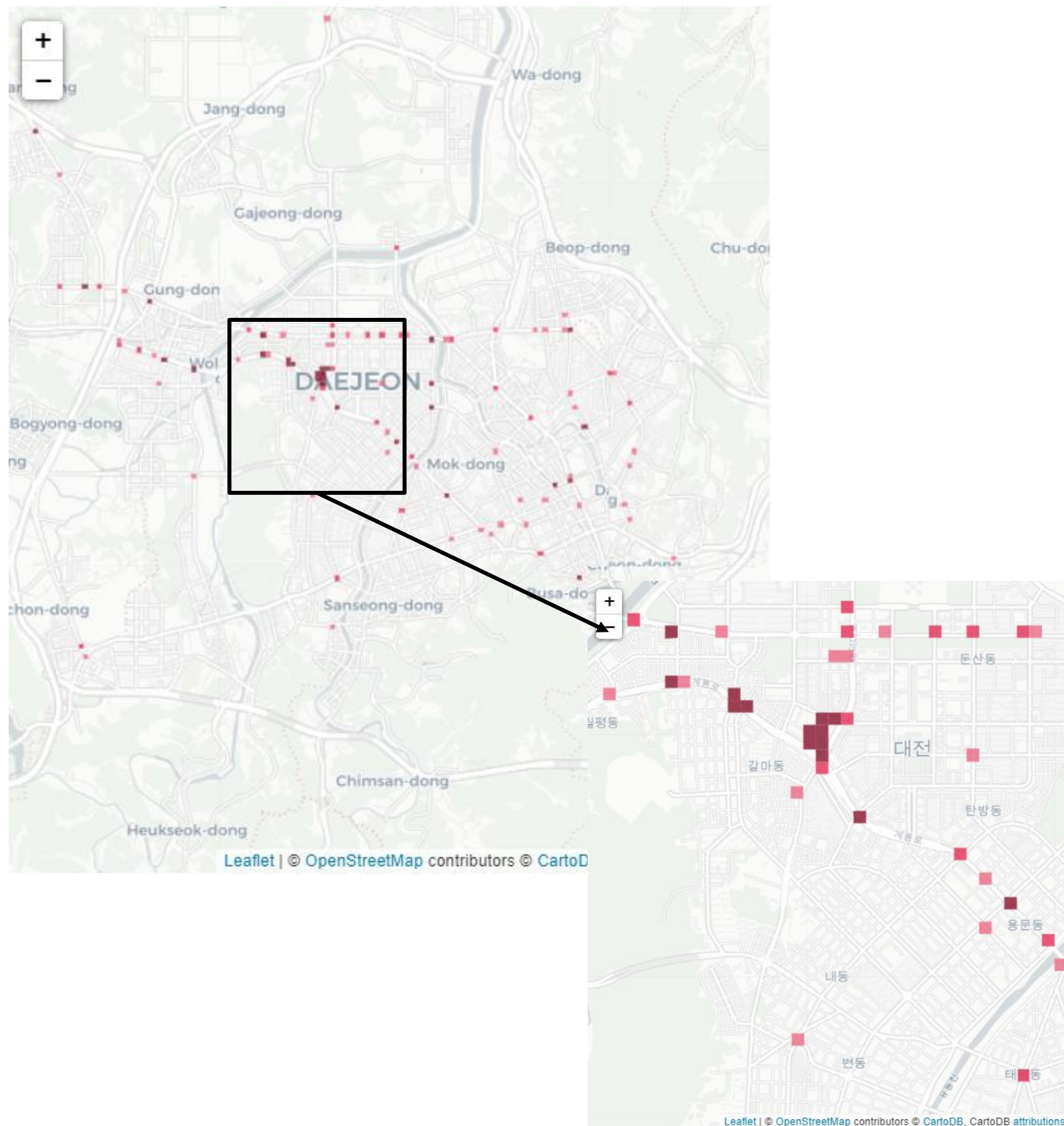
01 분석개요

✓ 분석목적

- 대전시 교통사고 예방을 위해 교통안전시설물 추가 설치를 계획함.
- 데이터 분석으로 우선 검토하여 교통사고 위험 지역 100개소를 선정.

✓ 분석방법

- 대전시 교통 사고 다발 지역 100개소 선정.
- 선정된 100개소를 기준으로 인구 분포, 차량 보유 수, 교통량 등을 분석하여 상관관계 파악.
- 100개소의 특성을 파악하여 설치가 필요한 적절한 교통안전시설물 파악 .

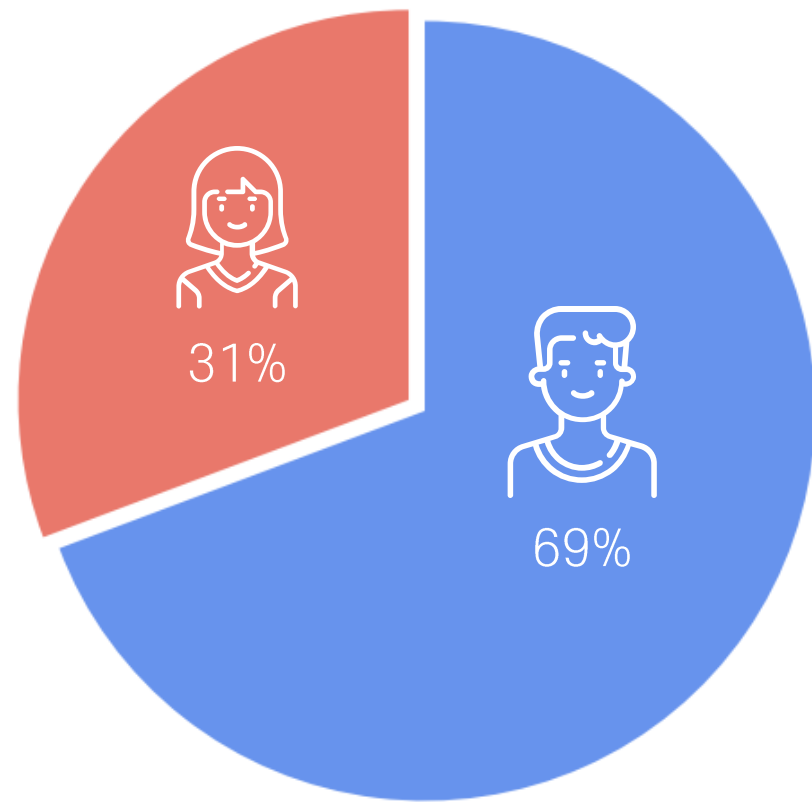


대전시 사고다발 지역 상위 100곳

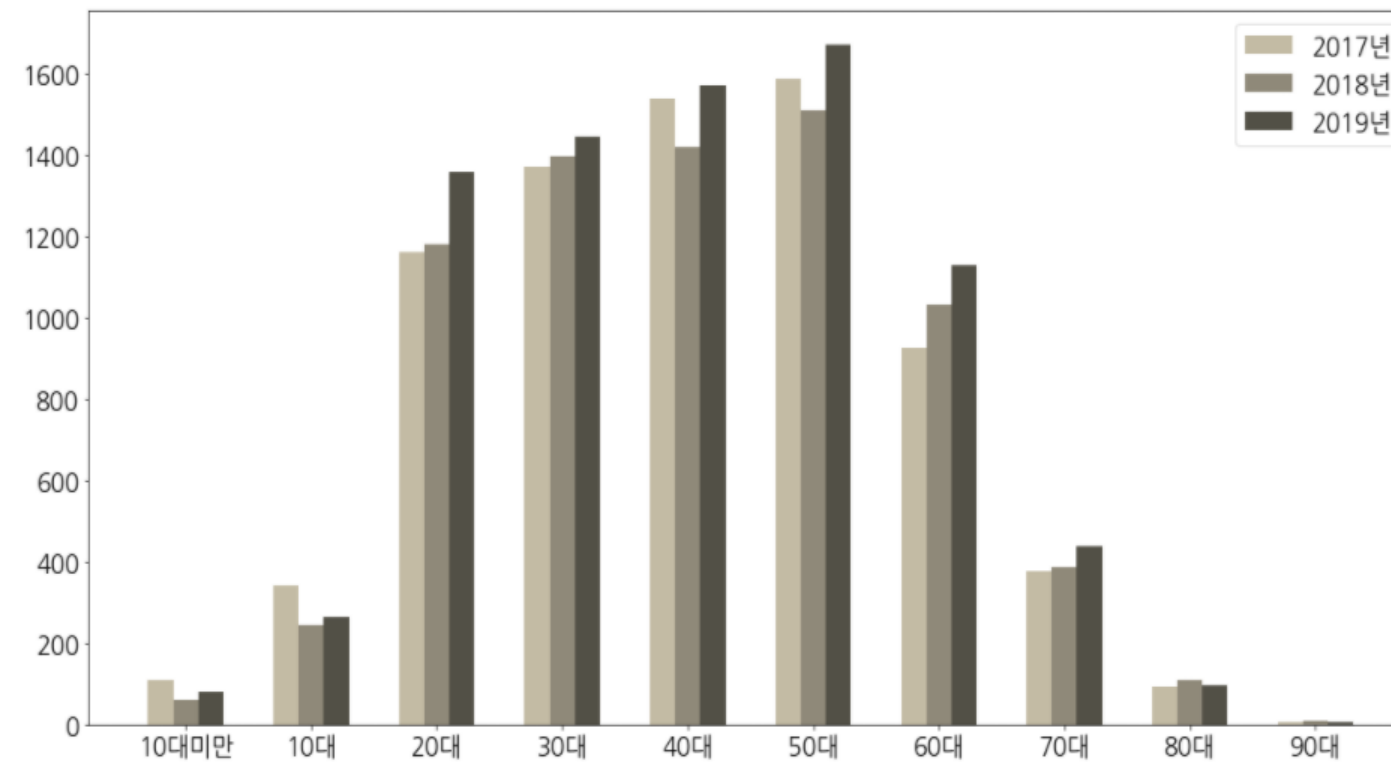
최근 3년 한곳에서 최대 63회의 사고 발생 지역 존재

2017년-2019년 교통사고 발생 자료를 분석한 결과 많은 도로에서 교통사고가 발생함을 알 수 있었으며, 가장 사고가 많이 나는 장소 100곳을 대전시 지도에 나타내 보았으며, 색이 진한 곳은 사고가 자주 발생하는 지역임을 알 수 있다.

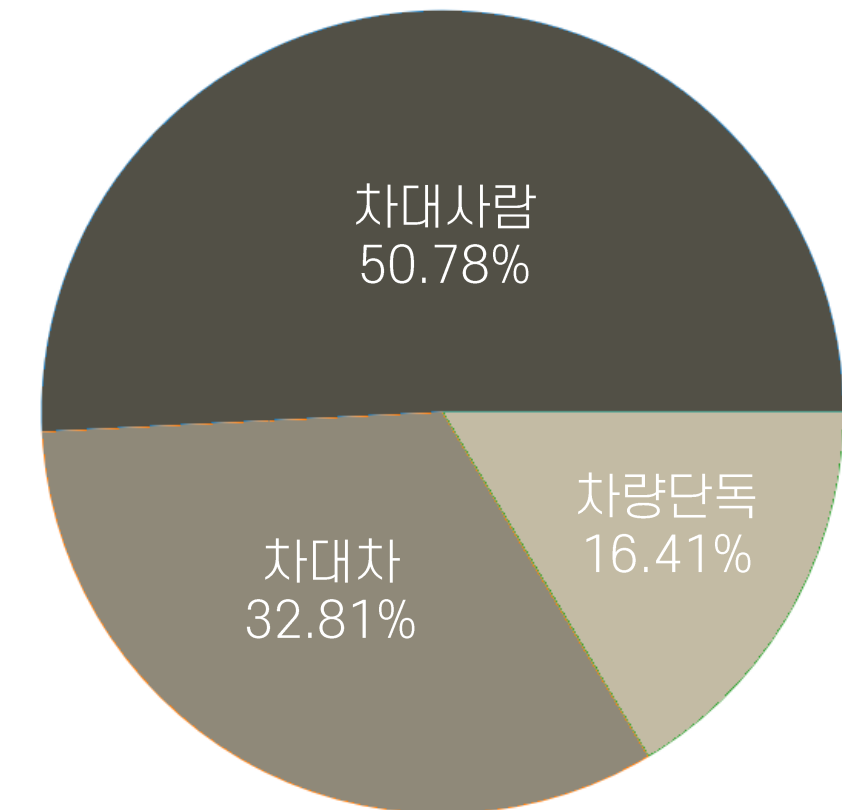
2017년-2019년도 교통사고 현황데이터를 통한 분석



대전시 교통사고 피해자 성별

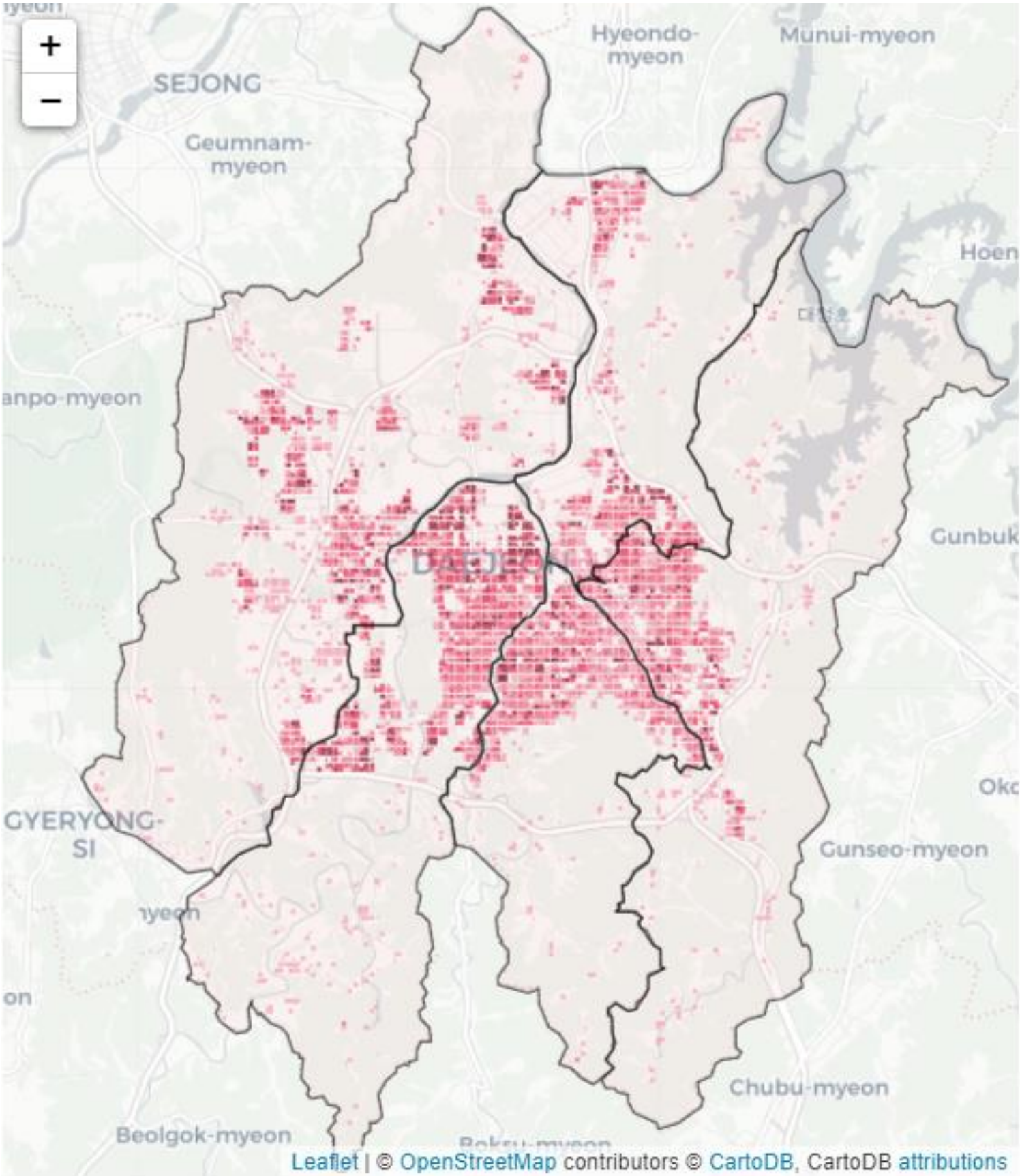


대전시 교통사고 연령대별 피해자



대전시 유형별 교통사고

교통사고 피해자는 남성의 비율이 69%로 여성에 비해 높았으며,
 ‘생산가능 인구’에 해당하는 30대-50대의 연령대의 사고피해자가 높게 나타나는 것을 볼 수 있다.
 교통사고의 유형을 분석한 결과 차대사람의 비율이 50.78%로 가장 높은 것을 확인할 수 있었다.



선택 이유

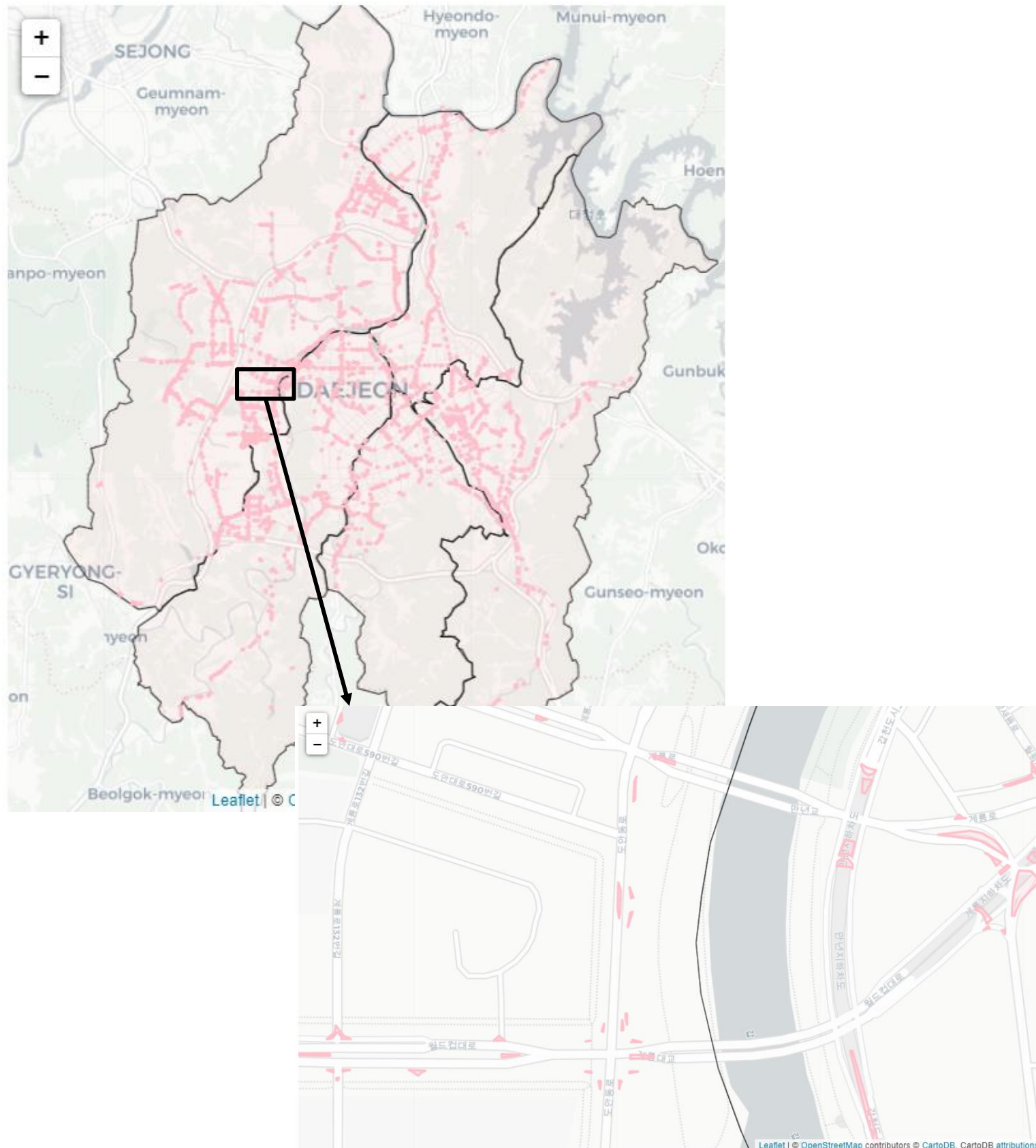
- 피해자 연령대별 분석 시 '10대미만', '10대'의 인구 데이터 수치가 적어 인구분포 분석에 영향을 미치지 않음.
- '총인구', '생산가능', '노령' 인구 데이터만을 사용.

전처리

- 'NaN'값은 실제지도와 비교한 결과 산, 도로 등 주거지역이 아니기 때문에 0으로 대체.

시각화

- 인구분포를 색의 진하기로 표현함.
->> 색이 진할수록 인구의 분포가 많음.



선정 이유

- 안전지대가 끝나는 부분 혹은 안전지대를 따라서 중앙분리대가 있는 경우가 많음.
- ‘중앙분리대’라는 데이터 대신 이 데이터를 사용.



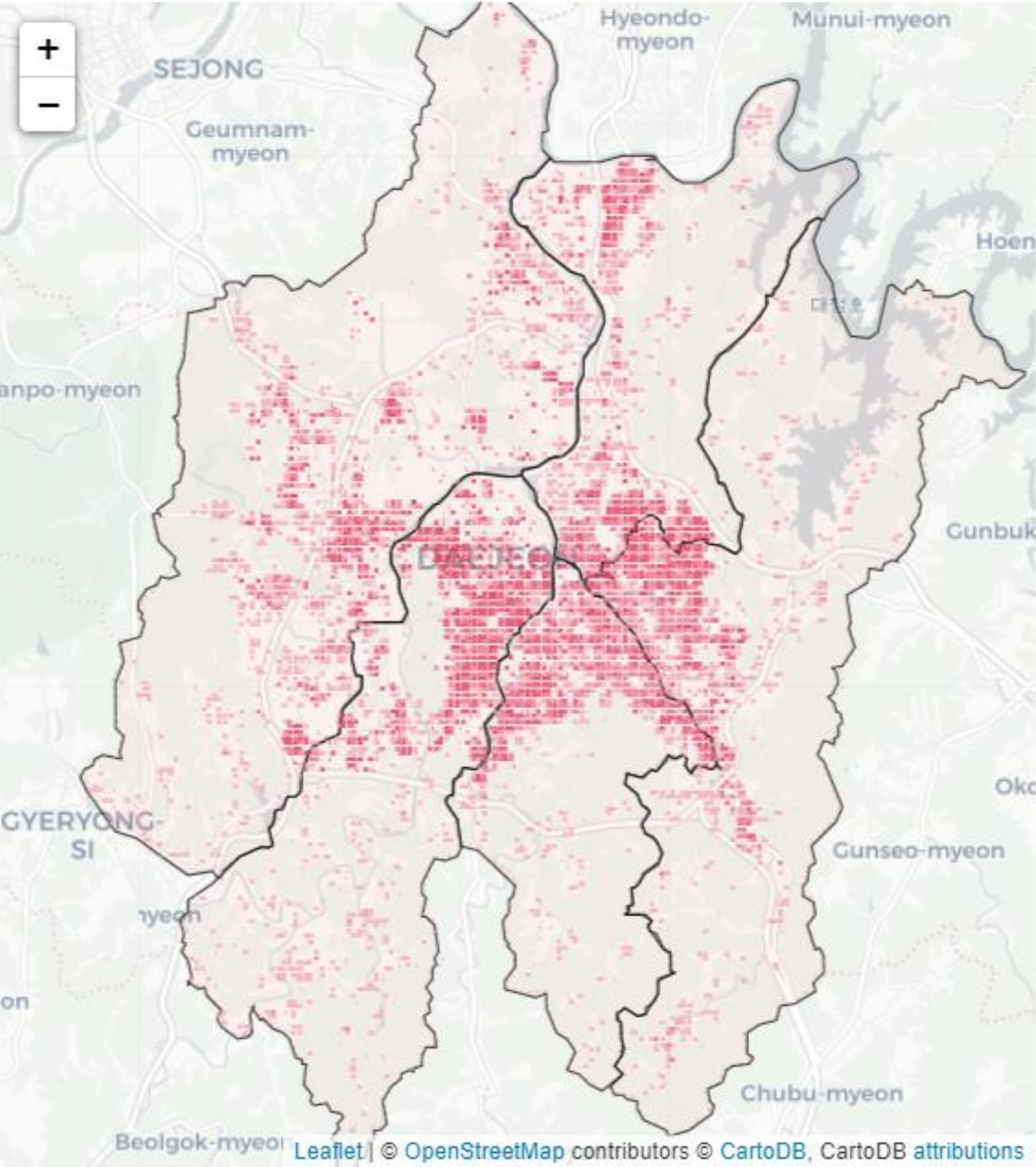
전처리

- 도로를 확대해보면 안전지대 모양이 전부 다름을 볼 수 있음.
->> ‘len’이 ‘NaN’인 경우가 많음
- MultiPolygon의 ‘area’를 구해 파생변수 ‘area’ 생성.
->>대전시 총 넓이인 539.8 km²에 맞춰 스케일 변환해줌.



시각화

- MultiPolygon 좌표 구현



선택 이유

- 차량등록수가 많은 지역은 차량의 유동이 많기 때문에 사고위험이 높다고 판단하여 이 데이터를 사용.

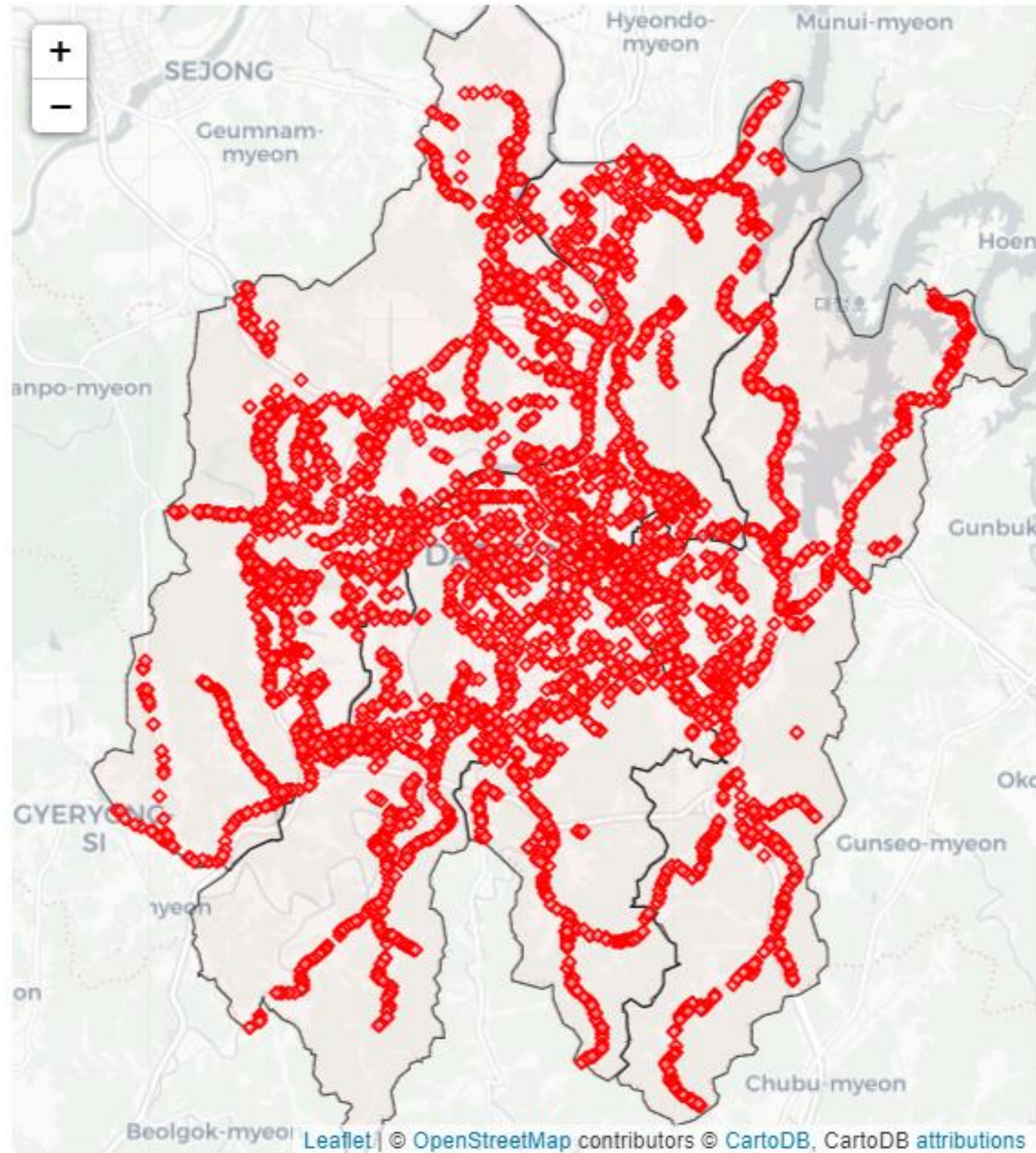
전처리

- 차량의 유동성을 보기위한 데이터로 반경 500m에 등록된 차량의 전체를 합하였음.

시각화

- 차량대수에 색의 진하기로 표현함.
 - 진한 빨강: 차량이 많음. 연한 빨강: 차량이 적음.

데이터명 : 30.대전광역시_차량등록현황_격자.geojson



선정 이유

- 지정된 위치에 따라 표지판의 수와 표지판의 중요도를 분석
- 중요도에 따라 높은 숫자를 부여하고, 표지판 개수에 따라 지정된 숫자를 합하여 최종 분석에 활용함.



전처리

- 지정된 위치에 주의, 규제, 지시, 보조의 표지 중 0~3 개의 표지판이 존재함.
- 주의, 규제, 지시, 보조 표지판을 중요도에 따라 각각 숫자 4, 3, 2, 1로 지정함.

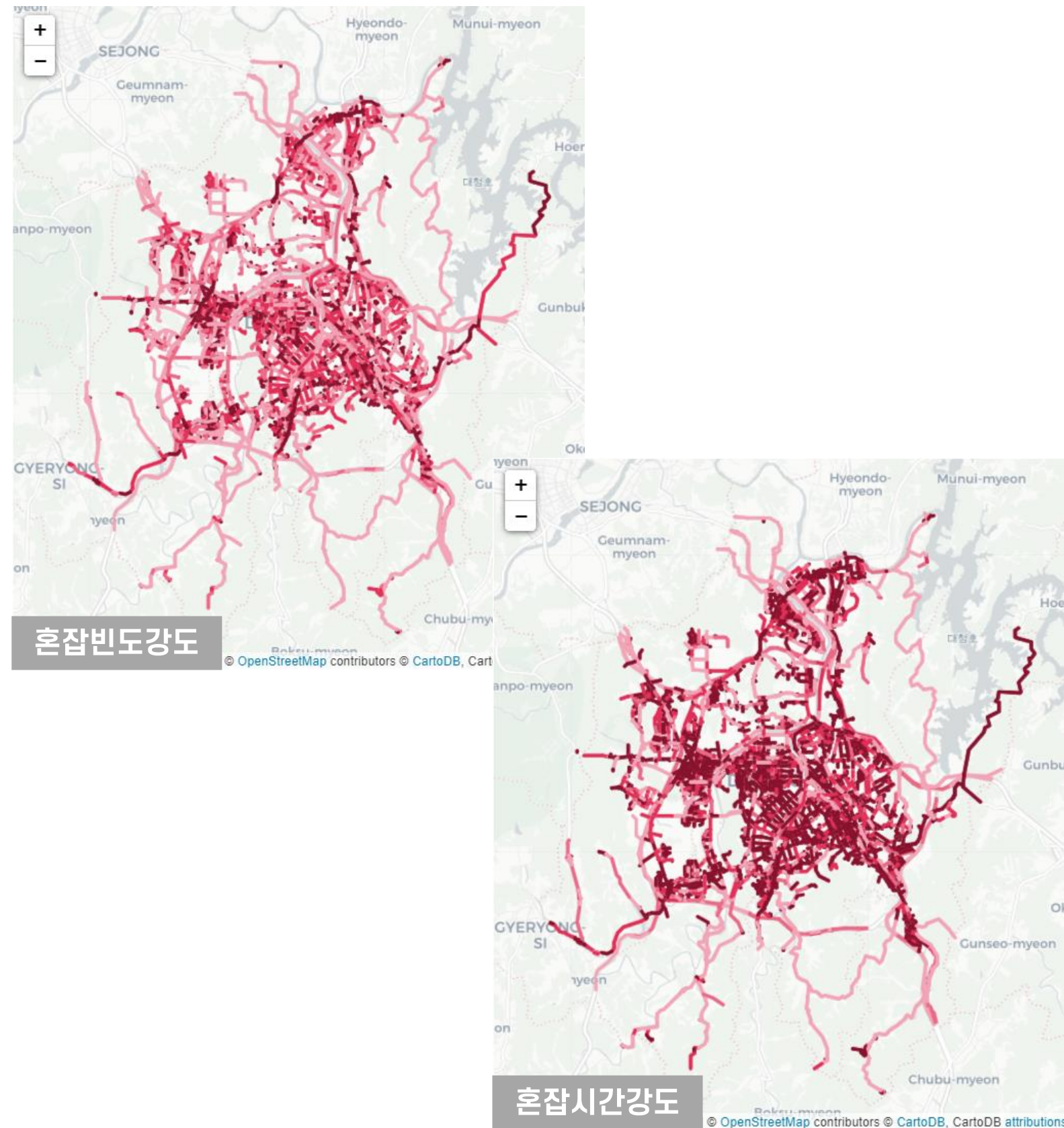


시각화

- 데이터양이 많기 때문에 주의, 규제, 지시, 보조 중 '주의' 표지만을 선택해서 시각화 함

데이터명 : 9.대전광역시_교통안전표지.geojson

분석과정 > 혼잡빈도강도, 혼잡시간강도



선택 이유

- 도로의 혼잡 정도와 혼잡 시간이 지속되면 사고의 위험과 큰 상관관계가 있다고 판단하여 이 데이터를 사용함.

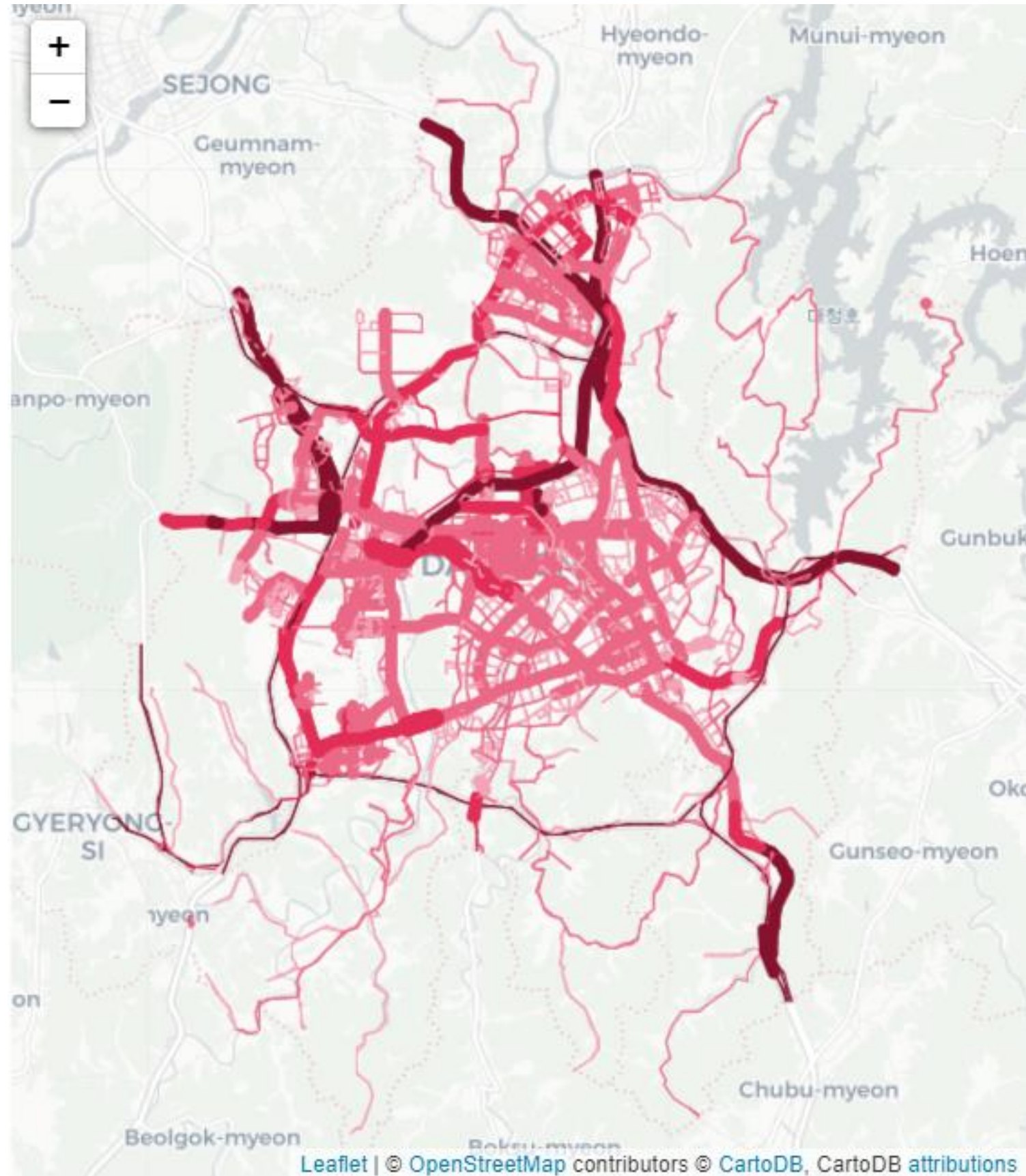
전처리

- 강도의 정도를 1-100까지로 나눠서 혼잡도가 높을수록 100에 가까운 큰 숫자를 부여함.

시각화

- 도로의 혼잡빈도(시간)강도에 따른 색 변화
->> 진한 빨강: 혼잡도가 높음, 연한 빨강: 혼잡도가 낮음

데이터명 : '21.대전광역시_평일_일별_혼잡빈도강도(2018).csv'
데이터명 : '22.대전광역시_평일_일별_혼잡시간강도(2018).csv'



선택 이유

- 도로의 제한속도와 도로의 차선 수는 교통사고 위험과 큰 상관관계가 있다고 판단하여 이 데이터를 사용함.

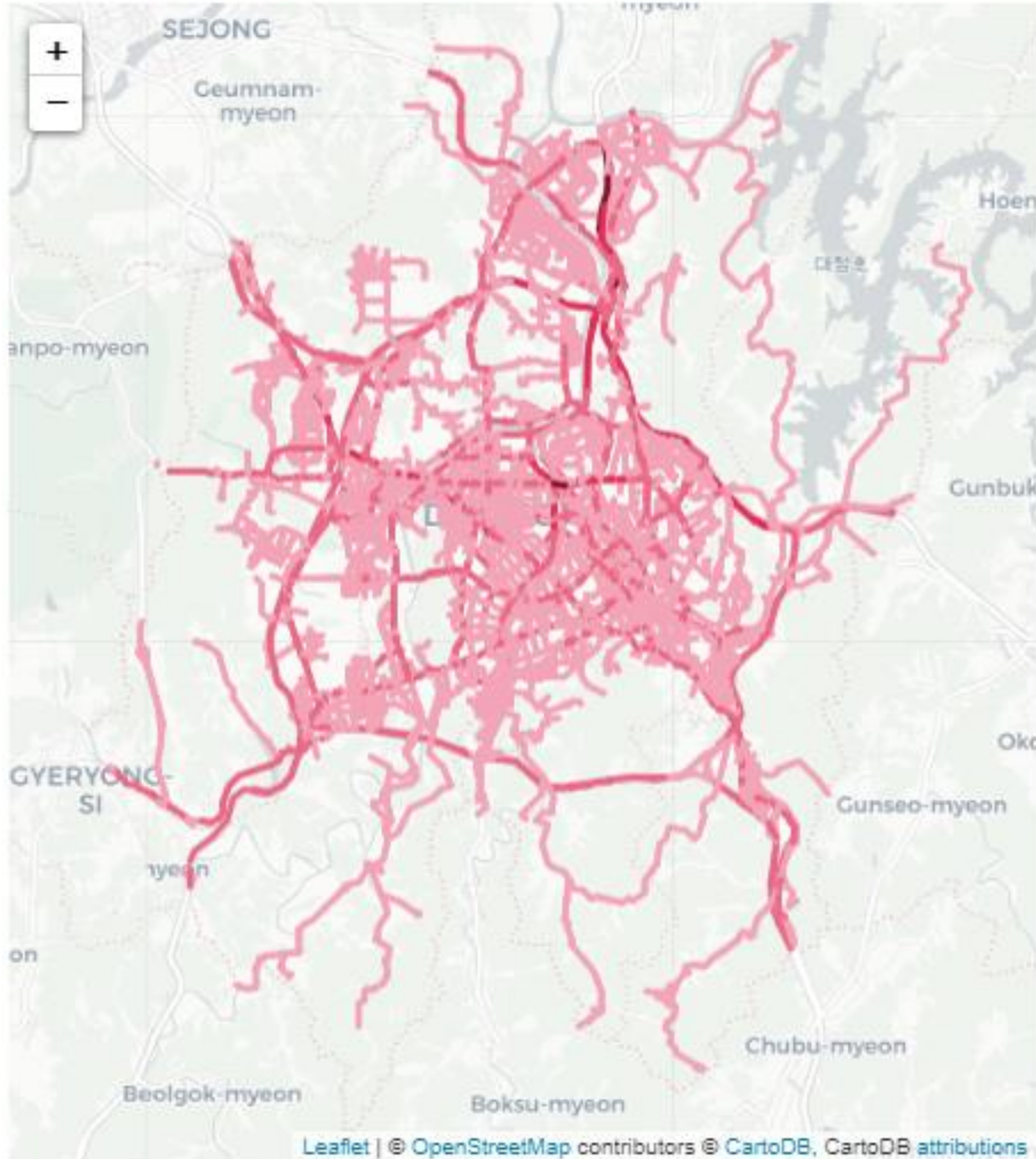
전처리

- 해당 구역에서 가장 가까운 도로의 제한속도, 차선 수, 도로 등급을 분석에 이용함

시각화

- 진한 빨강: 제한속도가 높음, 연한 빨강: 제한속도가 낮음
- 굵은 부분 : 차선수가 많음 얇은 부분 : 차선수가 적음

데이터명 : '7.대전광역시_도로속도표시.geojson'
데이터명 : '7.대전광역시_교통링크(2018).geojson'



선택 이유

- 하루동안 도로를 이용한 차량 수는 그 지점의 교통사고 발생율을 대변할 수 있다고 생각하여 이 데이터를 사용함.

전처리

- 0시부터 23시 & full time 데이터 중 full time 을 선택.
- 승용차, 화물차, 버스 3종의 총 교통량을 사용.

시각화

- 도로의 혼잡빈도(시간)강도에 따른 색 변화
->> 진한 빨강: 교통량 많음, 연한 빨강: 교통량 적음

데이터명 : '20.대전광역시_평일_일별_시간대별_추정교통량(2018).csv'

03 분석결과

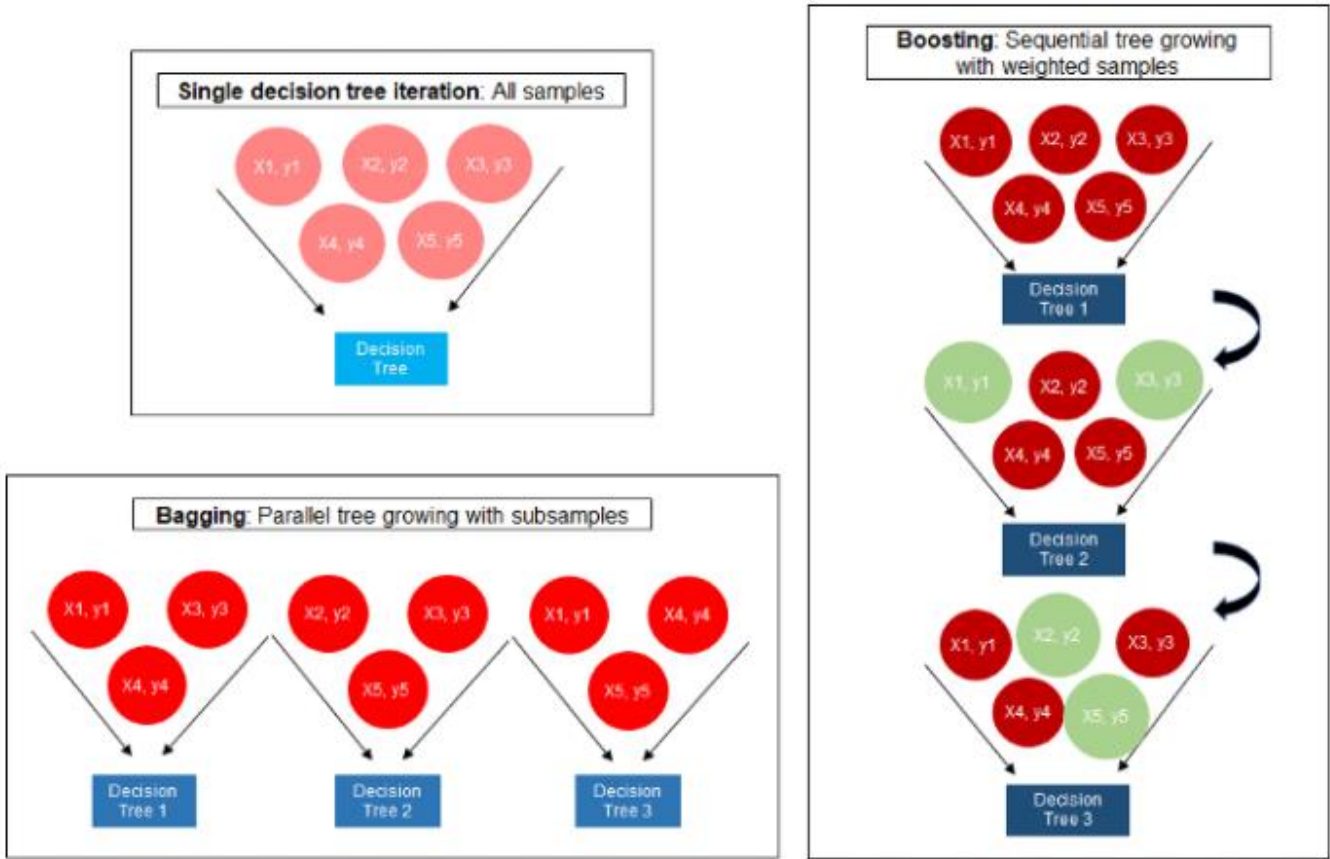
gid	보행등	차량등	안전지대	횡단보도_유무	속도표지판	안전표지판1st	안전표지판2nd	안전표지판3rd	교통CCTV_유무	차량노드	차로수	차로등급	속도제한	전일_전체교통량	혼잡빈도	혼잡시간	총인구수	고령인구수	생산가능인구수	차량등록현황	acci_cnt
16024 rows X 21 columns																					
다바 905151	1.0	0.0	0.0	2.0	0.0	2.0	0.0	0.0	0.0	0.0	2.0	104.0	60.0	NaN	NaN	NaN	11779.0	2253.0	8429.0	3808.0	4
다바 905166	0.0	0.0	0.0	0.0	0.0	11.0	0.0	0.0	0.0	0.0	3.0	104.0	60.0	10200.500000	32.040000	65.120	19190.0	2298.0	14919.0	6944.0	47
다바 888169	0.0	0.0	0.0	1.0	2.0	8.0	0.0	0.0	0.0	0.0	2.0	103.0	40.0	7211.666667	51.566667	76.300	19883.0	1918.0	16301.0	6825.0	7
다바 917174	2.0	4.0	0.0	3.0	0.0	7.0	0.0	0.0	0.0	0.0	3.0	104.0	60.0	13009.500000	38.095000	60.420	3547.0	833.0	2446.0	2463.0	2
다바 841191	4.0	6.0	0.0	3.0	0.0	13.0	0.0	0.0	0.0	1.0	2.0	104.0	40.0	NaN	NaN	NaN	12029.0	960.0	8777.0	5414.0	9

- 20개의 feature, 16024개 데이터를 train dataset으로 함
- Label을 사고발생횟수로 선택함

Feature 이름	Feature 설명	Feature 이름	Feature 설명
보행등	gid 반경 50m 내에 위치하는 보행등의 개수	차로수	gid 에서 가장 가까운 도로의 차로수
차량등	gid 반경 50m 내에 위치하는 차량등의 개수	차로등급	gid 에서 가장 가까운 도로의 차로등급 (예. 고속국도, 일반국도 등)
안전지대	gid 반경 50m 내에 위치하는 안전지대 면적의 합	속도제한	gid 에서 가장 가까운 도로의 최고속도
횡단보도	gid 반경 50m 내에 위치하는 횡단보도의 개수	전일_전체교통량	gid 반경 50m 내에 위치하는 도로들의 전일 교통량 평균
속도표지판	gid 반경 50m 내에 위치하는 속도표지판의 개수	혼잡빈도	gid 반경 50m 내에 위치하는 도로들의 혼잡강도 평균
안전표지판1st	gid 반경 50m 내에 위치하는 첫번째 안전표지판의 강도 점수 합	혼잡시간	gid 반경 50m 내에 위치하는 도로들의 혼잡시간 평균
안전표지판2nd	gid 반경 50m 내에 위치하는 두번째 안전표지판의 강도 점수 합	총인구수	gid 반경 500m 내에 거주하는 총인구수의 합
안전표지판3rd	gid 반경 50m 내에 위치하는 세번째 안전표지판의 강도 점수 합	고령인구수	gid 반경 500m 내에 거주하는 고령인구수의 합
교통CCTV	gid 반경 50m 내에 위치하는 교통 CCTV의 개수	생산가능인구수	gid 반경 500m 내에 거주하는 생산가능인구수의 합
차량노드	gid 반경 50m 내에 위치하는 교차로의 개수	차량등록현황	gid 반경 500m 내에 등록된 차량 개수의 합

- 앙상블 기법을 활용하면 과적합 감소시키며 예측 정확도를 높일 수 있음.
- 대표적인 앙상블 기법은 bagging과 boosting

비교	Bagging	Boosting
특징	병렬 앙상블 모델(각 모델은 서로 독립적)	연속 앙상블(이런 모델의 오류를 고려)
목적	Variance 감소	Bias 감소
적합한 상황	복잡한 모델(High variance, Low bias)	Low variance, High bias 모델
대표 알고리즘	Random Forest	Gradient Boosting, AdaBoost
sampling	Random Sampling	Random Sampling with weight on error



- Bagging을 이용한 대표적 모델은 Randomforestregressor , Boosting을 이용한 대표적인 방법은 XGBoost
- 일반적으로 XGBoost는 대량의 데이터에서 더 효율적이므로 Randomforestregressor 을 모델로 채택.

Model : RandomForestRegressor Loss : 0.29448711111111106

(예상 사고 횟수와 실제 사고 횟수 차이 제곱의 최댓값)

Parameter	rf_random.best_params
n_estimators	125
min_samples_split	5
min_samples_leaf	1
max_features	'sqrt'
max_depth	90
bootstrap	False

- Model : xgboost.XGBRegressor Loss : 67.85617371579065 다른 모델을 적용한 결과 Loss값의 차이가 확연함을 확인할 수 있었다.

	longitude	latitude	acci_result
0	127.390068	36.358143	19.592000
1	127.371330	36.351513	13.453333
2	127.461899	36.318138	13.384000
3	127.455559	36.354698	11.014667
4	127.430031	36.327591	9.704000
...
95	127.352936	36.285900	1.293333
96	127.478752	36.382938	1.292000
97	127.402924	36.366978	1.270667
98	127.341385	36.414001	1.242667
99	127.332680	36.249329	1.069333

100 rows × 3 columns

✓ 최종 파라미터 적용 결과

지도 상의 도로 위에 대량(5721개)의 임의의 점 생성

→ 학습된 모델로 예상사고횟수 도출

→ 해당 예상사고횟수를 가중치로 두고 K means 군집화(n_clusters=100) 진행

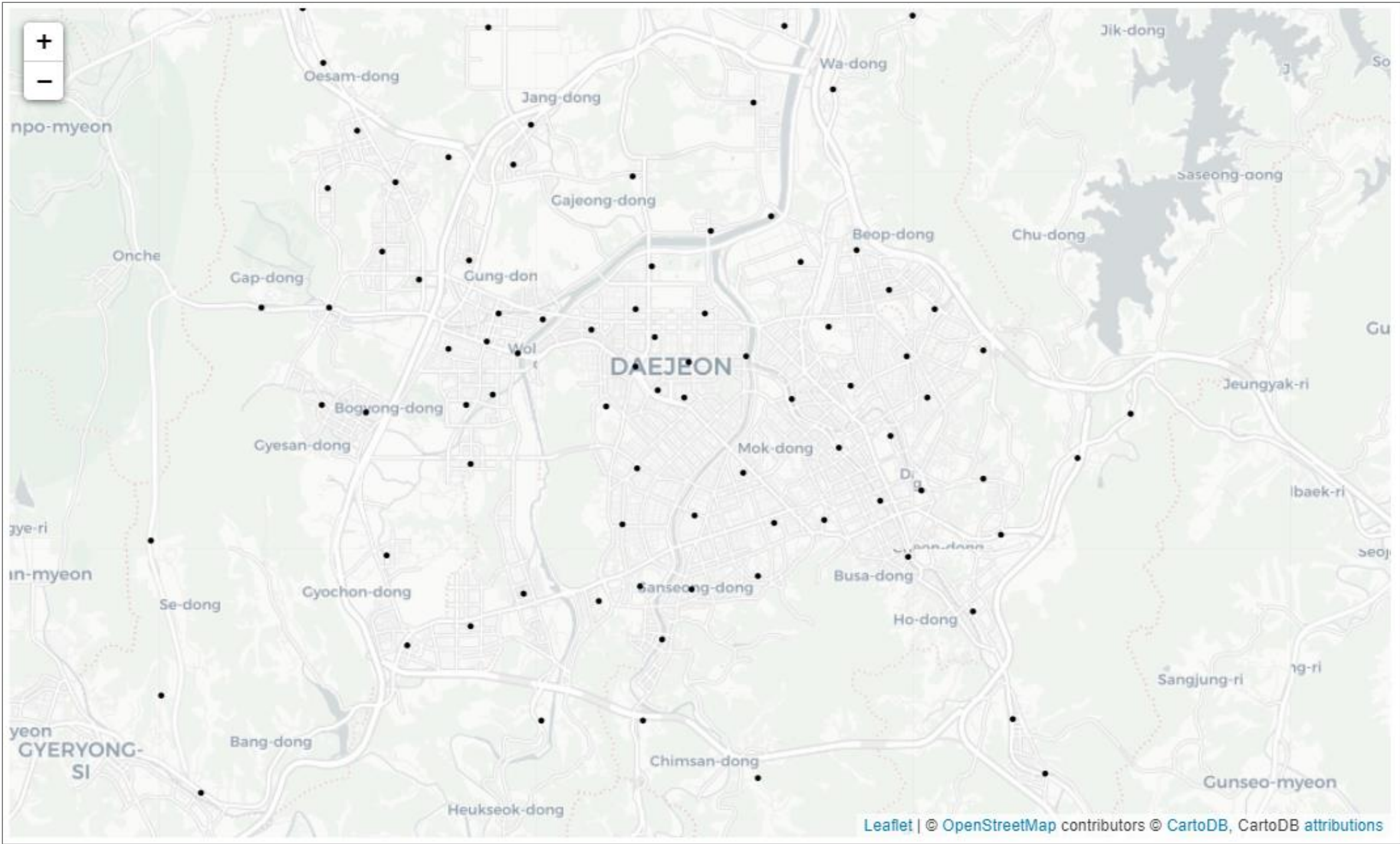
→ 군집화 결과 좌표를 가까운 도로로 이동

→ 이동시킨 최종 100 곳의 위도, 경도

✓ K means 군집화

	longitude	latitude	acci_result
0	127.390068	36.358143	19.592000
1	127.371330	36.351513	13.453333
2	127.461899	36.318138	13.384000
3	127.455559	36.354698	11.014667
4	127.430031	36.327591	9.704000
...
95	127.352936	36.285900	1.293333
96	127.478752	36.382938	1.292000
97	127.402924	36.366978	1.270667
98	127.341385	36.414001	1.242667
99	127.332680	36.249329	1.069333

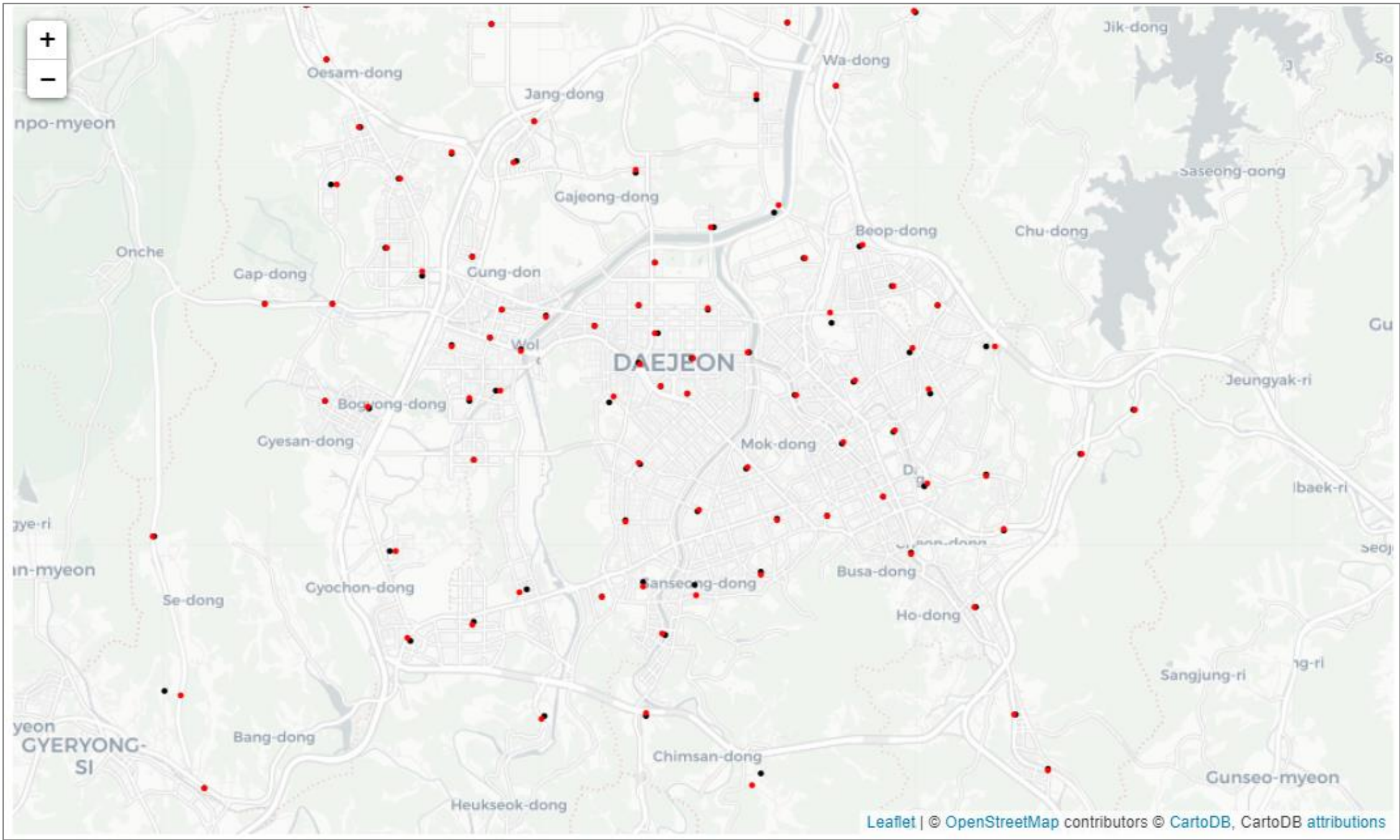
100 rows x 3 columns



✓ 군집화한 점을 가까운 도로로 이동

	longitude	latitude	acci_result
0	127.390068	36.358143	19.592000
1	127.371330	36.351513	13.453333
2	127.461899	36.318138	13.384000
3	127.455559	36.354698	11.014667
4	127.430031	36.327591	9.704000
...
95	127.352936	36.285900	1.293333
96	127.478752	36.382938	1.292000
97	127.402924	36.366978	1.270667
98	127.341385	36.414001	1.242667
99	127.332680	36.249329	1.069333

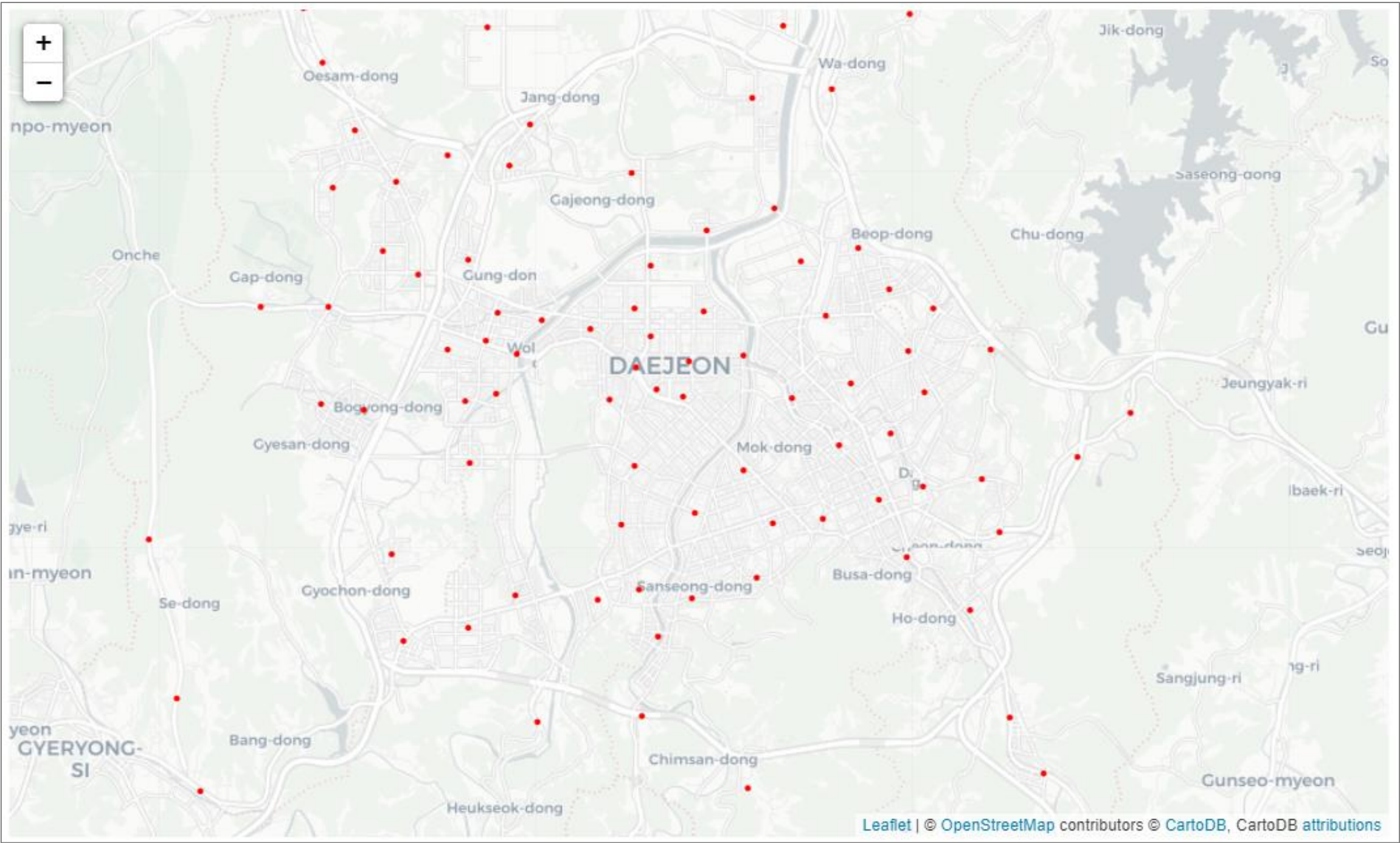
100 rows x 3 columns



✓ 최종 선정한 100곳

	longitude	latitude	acci_result
0	127.390068	36.358143	19.592000
1	127.371330	36.351513	13.453333
2	127.461899	36.318138	13.384000
3	127.455559	36.354698	11.014667
4	127.430031	36.327591	9.704000
...
95	127.352936	36.285900	1.293333
96	127.478752	36.382938	1.292000
97	127.402924	36.366978	1.270667
98	127.341385	36.414001	1.242667
99	127.332680	36.249329	1.069333

100 rows × 3 columns



✓ 최종 선정한 100곳

위험순위	시설명	주소지	X좌표(경도)	Y좌표(위도)	반경범위
0	대전광역시 서구 둔산동	다바889168	127.376877	36.349129	반경 50m
1	대전광역시 동구 용전동	다바944169	127.438179	36.349556	반경 50m
2	대전광역시 유성구 송강동	다바905267	127.395088	36.438694	반경 50m
3	대전광역시 서구 둔산동	다바891176	127.379295	36.356415	반경 50m
4	대전광역시 서구 갈마동	다바883162	127.369980	36.343754	반경 50m
5	대전광역시 서구 정림동	다바883121	127.370033	36.306988	반경 50m
6	NaN	다바960174	127.455444	36.354637	반경 50m
7	대전광역시 동구 삼성동	다바941156	127.435249	36.338375	반경 50m
8	대전광역시 유성구 용산동	다바911249	127.400841	36.422382	반경 50m
9	대전광역시 서구 월평동	다바864171	127.348892	36.351593	반경 50m

최종 선정한 100곳의 위경도를 이용하여 gid 추출 후 '1.대전광역시_교통사고내역(2017~2019).csv'와 연결하여 시설명 도출

감사합니다.