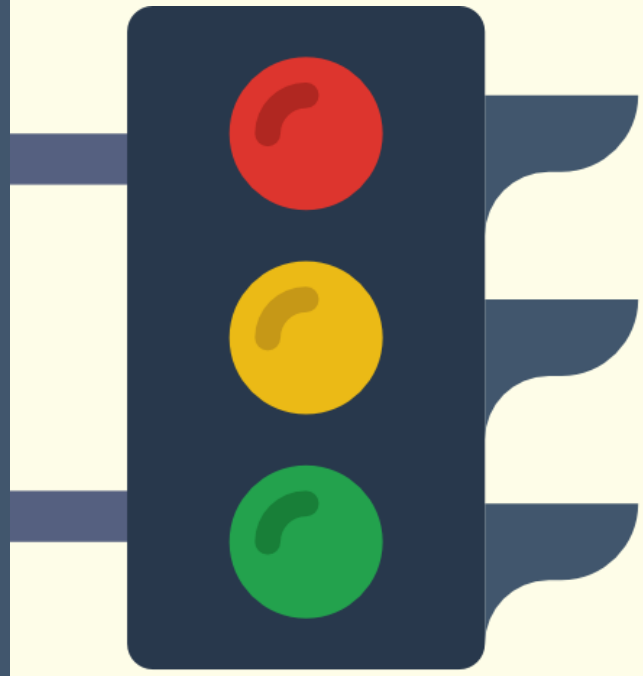


“포아송 회귀모형”을 기반으로 교통사고 위험지역 도출하기

팀명 : 백년보장무사고

진행과정

목차

content

Ⅰ 서론

i . 분석의 방향성 소개

Ⅱ 본론

i . 포아송회귀분석

a. 회귀분석 데이터 생성

1. Y(종속변수) 생성

2. X(독립변수) 생성

b. 포아송회귀모형 적합

c. 결과해석

ii . 시각화

Ⅲ 결론

i . 시각화 해석 및 위험지역 도출

Ⅰ 서론

a. 분석의 방향성 소개

a. 분석방법 개요 - 포아송 회귀 분석 소개

공모대회에서 제공하는 자료를 바탕으로 진행한 연구과제는
교통사고 발생양상과 관련된 여러 잠재요인들을 자료에서 선별하고
해당 요인들의 효과를 통계적으로 추론하는 과정을 거쳤습니다.

이를 위해 '시군구' 별로 자료를 count하고,
계수형 자료(count data)에 최적화된 포아송 회귀모형을
적합시켰습니다. 모형 적합 이후 다중공선성 검진의 절차를
수행하여 최고의 성능을 보이는 모형을 선택했습니다.

궁극적으로는 최종모형으로부터 교통사고 발생 확률을 예측하기 위한
주요 요인을 탐색하여 이를 해석하는 데 있습니다.

Ⅱ 본론

ⅰ. 포아송 회귀분석

a. 회귀분석 데이터 생성

1. Y(종속변수) 생성
2. X(독립변수) 생성

b. 포아송 회귀모형 적합

c. 결과해석

ⅱ. 시각화

2. 본론

i . 포아송 회귀분석

✓ 사전 작업
제공되는 자료 속 '시군구 ' column을 기준으로
종속변수, 독립변수를 모두 계수형 자료로
변환하기 위해 필요한 기준 데이터 셋 만들기.

✓ 총 5개의 동으로 이루어진 대전 속에서 각각 대응
되는 동을 매칭시켜 완성 함.

✓ 기준 데이터 셋
: 법정동 기준 총 177개의 ~구 ~동 형태

✓ 제출한 코드 : data_tot 의 '시군구' column

* 법정동 기준으로 만든 이유
: 행정기준은 총 79개 동으로 포아송 회귀분석을 진행하는데
sample size가 작다 판단되어 법정기준을 선택함

	시군구
0	대덕구 문평동
1	동구 신촌동
2	유성구 어은동
3	대덕구 와동
4	유성구 복용동
...	...
172	대덕구 오정동
173	동구 소제동
174	중구 문화동
175	동구 신상동
176	동구 신흥동

177개

2. 본론

i . 포아송 회귀분석

- 회귀분석 데이터 생성
- Y(종속변수) 사고건수

✓ 사전작업으로 만든 '시군구' 기준으로
제공된 교통사고건수 내역
(1.대전광역시_교통사고내역(2017~2019).csv)을
시군구 기준으로 count한 후, y(종속변수)로 만듦.

✓ 제출한 코드 : data_tot

* 마지막에 시군구 별 인구자료를 외부로 부터 찾아
사고건수를 시군구 별 인구로 나눈 비율로 나타냄.

종속변수

	시군구	사고 건수
0	대덕구 문평동	59.0
1	동구 신촌동	NaN
2	유성구 어은동	51.0
3	대덕구 와동	44.0
4	유성구 복용동	42.0
...	...	...
172	대덕구 오정동	571.0
173	동구 소제동	76.0
174	중구 문화동	339.0
175	동구 신상동	25.0
176	동구 신흥동	94.0

177 rows × 2 columns

2. 본론

i. 포아송 회귀분석

- 회귀분석 데이터 생성
- x(독립변수) 생성

✓ 앞서 만든 데이터셋에 독립변수(x)를 이어 붙임.
✓ '시군구' 기준으로 6가지 자료를 count

✓ 6가지 자료
① 3.대전광역시_신호등(보행등).geojson
② 10.대전광역시_교통CCTV.geojson
③ 6.대전광역시_횡단보도.geojson
④ 7.대전광역시_도로속도표시.geojson
⑤ 4.대전광역시_신호등(차량등).geojson
⑥ 9.대전광역시_교통안전표지.geojson

Geomtry 변수가 point로 Count가 가능한변수들

data_tot에서 CCTV결측을 0으로 대체하고
나머지 변수들의 결측치가 몇 개인지 확인한 결과

시군구	0
사고 건수	15
신호등개수	44
cctv개수	0
횡단보도 개수	17
도로속도 표시	38
차량등개수	46
교통안전표지	11

✓ 결측치 대체
: 이어 붙이는 과정에서 기준 '시군구'
에 해당하는 데이터가 없는 경우도 발생.

✓ CCTV 변수 결측
: 전부 0으로 대체
CCTV에 결측이 많아서 찾아본 결과,
실제로도 CCTV가 많이 없다는 사실을
(스마트도시홈페이지 - cctv관제현황)를 통해 알게 됨.

✓ CCTV제외한 나머지 변수 결측
: KNN 방법을 통해 결측치를 대체

2. 본론

i . 포아송 회귀분석

✓ 앞서 만든 데이터셋(data_tot)에 6가지 독립변수(x)를 이어 붙이고 결측치 처리 후 '시군구' 열 없앤 데이터 셋
✓ 제출한 코드 : df22

6가지 독립변수

	사고 건수	신호등개수	cctv개수	횡단보도 개수	도로속도 표시	차량등개수	교통안전표지
0	59.0	28.0	1.0	43.0	54.0	41.0	162.0
1	1.6	5.6	0.0	1.0	1.0	10.0	18.0
2	51.0	19.0	0.0	29.0	2.0	19.0	47.0
3	44.0	14.0	0.0	41.0	30.0	31.0	226.0
4	42.0	46.0	0.0	110.0	8.0	83.0	453.0
...	...	...	...	...	...	...	...
172	571.0	102.0	9.0	278.0	126.0	197.0	881.0
173	76.0	33.0	0.0	41.0	17.0	64.0	126.0
174	339.0	153.0	2.0	222.0	56.0	209.0	574.0
175	25.0	6.0	0.0	4.0	20.0	15.0	97.0
176	94.0	47.0	0.0	48.0	3.0	66.0	115.0

2. 본론

ⅰ. 포아송 회귀분석

- Y(종속변수) 보완
  : 사고건수(y)를 각 지역당 인구로 나눈 비율로 환산하는 과정

✓ 제출한 코드 : data_peo

-〉 data_peo는 구,동에 맞는 인구를 외부데이터에서 찾아 엑셀파일로 정리 한 후 불러온 데이터 셋임.

-〉 정보가 없는 지역을 제외하고 총 145개의 인구정보를 data_peo에 담았음.

data_peo

	시군구	인구수
0	동구 가양1동	13927
1	동구 가양2동	18360
2	동구 가오동	12858
3	동구 구도동	92
4	동구 낭월동	10667
...	...	...
140	대덕구 비래동	6940
141	대덕구 송촌동	28591
142	대덕구 중리동	9877
143	대덕구 법동	26971
144	대덕구 석봉동	14417

✓ 참조한 외부데이터 사이트

1. 유성구 인구 : 공공데이터 포탈 - 데이터 찾기 - 대전광역시_유성구_인구통계 현황
2. 동구 인구 : 대전광역시 동구청 홈페이지 - 동구소개 - 통계정보 - 인구통계 - 주민등록(법정동별) 인구통계(2021.3.31)
3. 중구 인구 : 공공데이터 포탈 - 데이터 찾기 - 검색 : 대전광역시 중구 인구 - 대전광역시 중구_인구통계
4. 서구 인구 : 대전광역시 서구청 홈페이지 - YES 서구 - 우리구 소개 - 서구 통계 정보 - 2021년 3월 주민등록 인구현황 - 2021년 3월 인구현황.xlsx
5. 대덕구 인구 : 대전광역시 대덕구청 홈페이지 - 대덕구 소개 - 대덕 통계 - 주민등록 현황 - 인구형황 (2021.3)

2. 본론

i. 포아송 회귀분석 최종 데이터 셋

종속변수

	사고 건수	신호등개수	cctv개수	횡단보도개수	도로속도표시	차량등개수	교통안전표지	시군구	인구수	인구당_사고건수
1	1.6	5.6	0.0	1.0	1.0	10.0	18.0	동구 신촌동	50.0	0.032000
2	51.0	19.0	0.0	29.0	2.0	19.0	47.0	유성구 어은동	10395.0	0.004906
4	42.0	46.0	0.0	110.0	8.0	83.0	453.0	유성구 복용동	1262.0	0.033281
6	325.0	103.0	2.0	143.0	97.0	166.0	480.0	유성구 노은동	15589.0	0.020848
7	10.0	2.0	0.0	6.0	6.0	3.0	44.0	동구 추동	382.0	0.026178
...	...	...	...	...	...	...	...	...	...	...
172	571.0	102.0	9.0	278.0	126.0	197.0	881.0	대덕구 오정동	15125.0	0.037752
173	76.0	33.0	0.0	41.0	17.0	64.0	126.0	동구 소제동	2039.0	0.037273
174	339.0	153.0	2.0	222.0	56.0	209.0	574.0	중구 문화동	35913.0	0.009439
175	25.0	6.0	0.0	4.0	20.0	15.0	97.0	동구 신상동	203.0	0.123153
176	94.0	47.0	0.0	48.0	3.0	66.0	115.0	동구 신흥동	4824.0	0.019486

data_peo와 data_real_tot(df_22에서 '시군구' 열을 다시 붙인 데이터 셋) 를 '시군구'로 merge한 후 인구 정보를 고려하여 사고 건수를 파악하기 위해 인구당_사고건수 변수를 생성하여 data_real_tot2 라는 최종 데이터셋을 생성 함.

2. 본론

ⅰ. 포아송 회귀분석

- 포아송 회귀모형 적합 과정

Step1. 생성한 6개의 독립변수 모두 포함하여 회귀모형 적합
-> 유의한 독립변수가 2개 존재하지만 신호등개수와 차량등개수 사이의 다중공선성 문제가 크게 발생

Step2. 차량등개수를 제거한 5개 독립변수 포함하여 회귀모형 적합
-> 유의한 독립변수가 1개 존재하나 횡단보도 개수와 교통안전표시 사이의 다중공선성 문제 발생

Step3. 차량등개수와 횡단보도 개수를 제외한 4개의 독립변수 포함하여 회귀모형 적합
-> 신호등개수와 cctv개수 회귀계수가 유의하며 vif 지수가 모두 10이하로 다중공선성 하여 모형으로 step3모형을 채택

2. 본론

i. 포아송 회귀분석

- 적합한 회귀모형 해석

Dep. Variable:	인구당_사고건수	**No. Observations:**	139
Model:	GLM	**Df Residuals:**	135
Model Family:	Poisson	**Df Model:**	3
Link Function:	log	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-48.381
Date:	Thu, 01 Apr 2021	**Deviance:**	50.098
Time:	12:13:43	**Pearson chi2:**	1.10e+07
No. Iterations:	7		
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
신호등개수	-0.0427	0.030	-1.436	0.151	-0.101	0.016
cctv개수	0.5931	0.280	2.120	0.034	0.045	1.141
도로속도표시	0.0090	0.026	0.349	0.727	-0.042	0.060
교통안전표지	-0.0126	0.006	-2.126	0.034	-0.024	-0.001

모형3의 다중공선성 결과

	VIF Factor	features
0	2.061090	Intercept
1	6.908633	신호등개수
2	1.804998	cctv개수
3	2.404609	도로속도표시
4	7.412085	교통안전표지

3) 회귀분석 결과

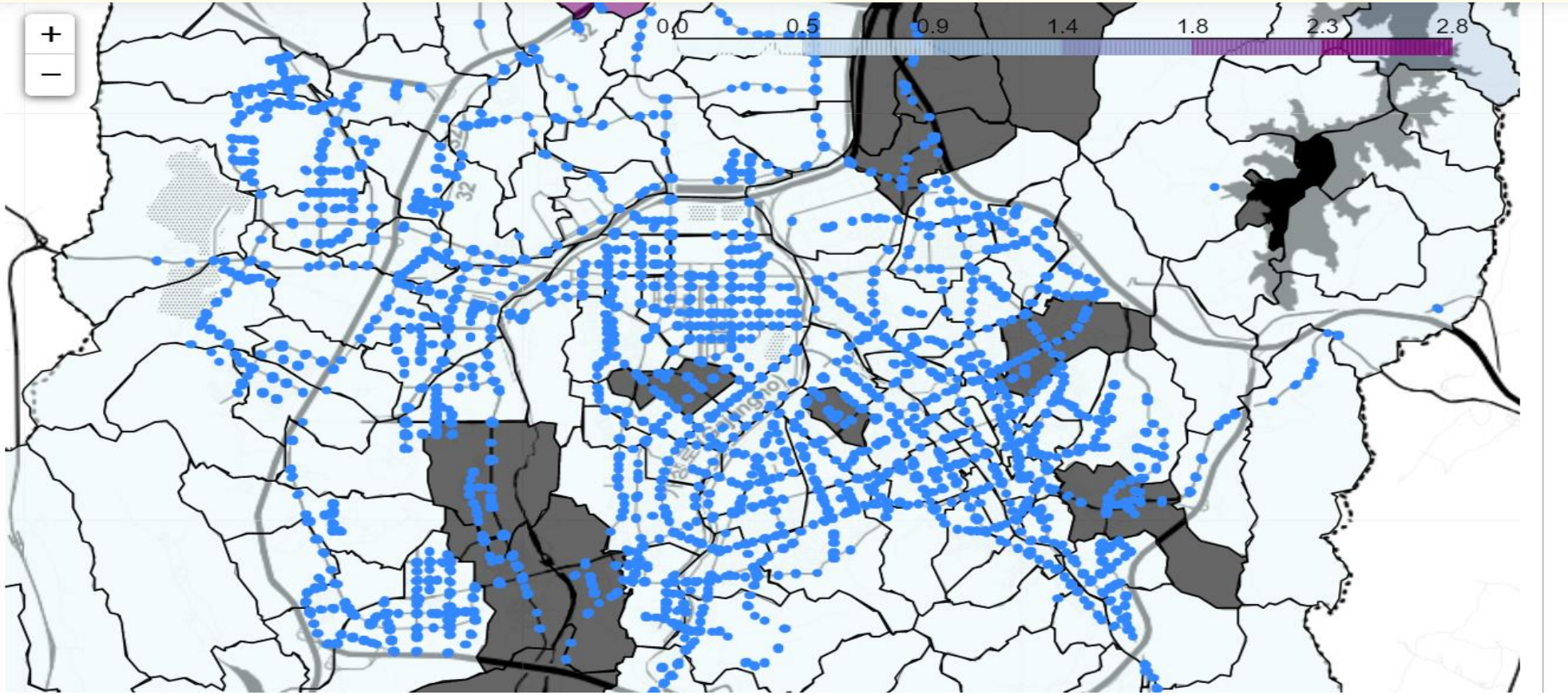
앞서 신호등개수와 차량등개수, 횡단보도개수와 교통안전표지 사이의 다중공선성 존재함을 확인함 따라서 차량등을 제거하고, 횡단보도와 교통안전표지를 각각 제거한 결과를 비교하여 최종적으로 신호등개수, cctv개수, 도로속도표시, 교통안전표지를 독립변수로 채택함.

CCTV개수와 교통안전표지는 인구당_사고건수에 영향을 주는 유의한 변수임을 알 수 있음.

2. 본론

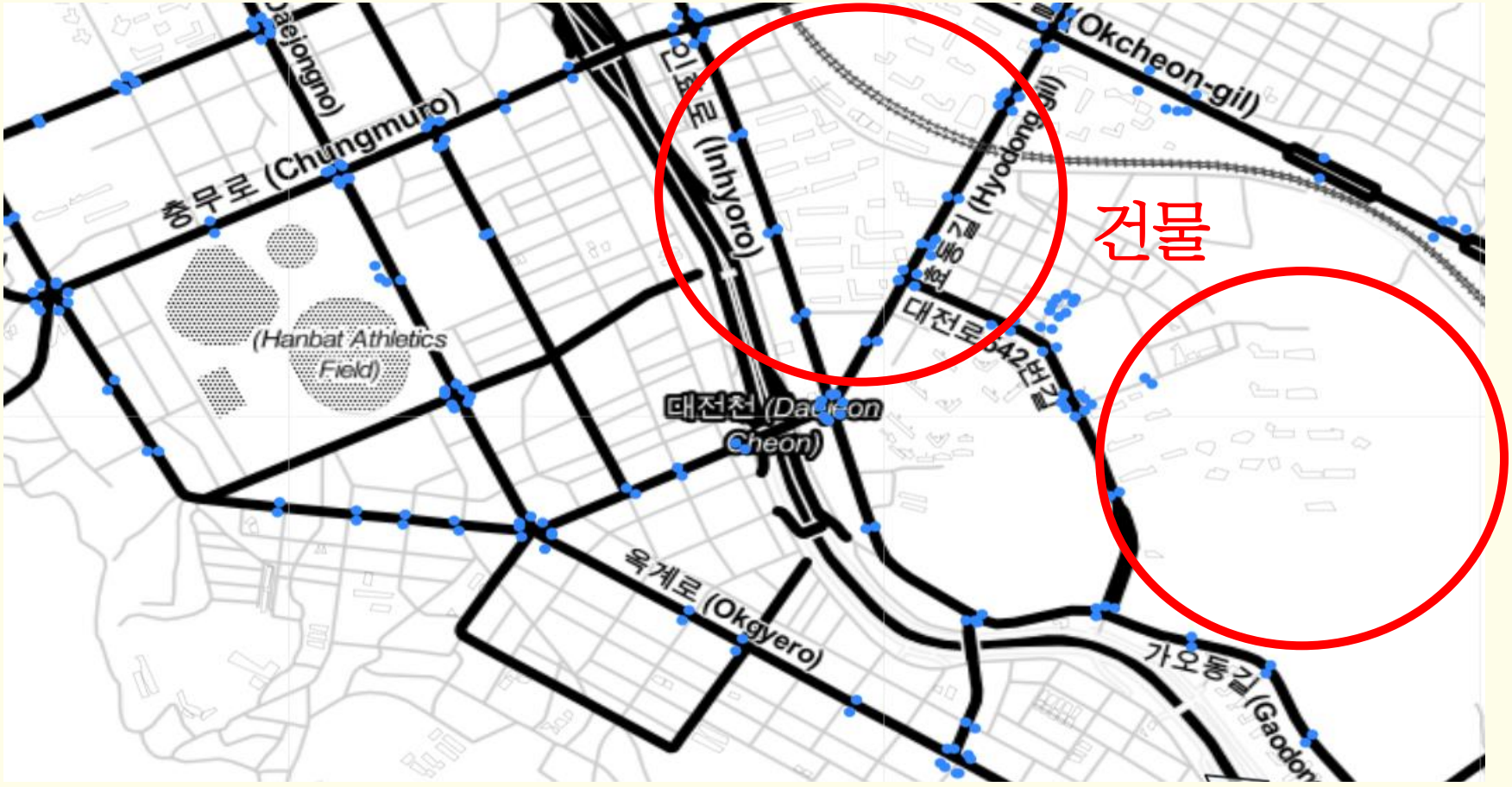
ii. 시각화

대전시 '시군구' 를 map에 나타내는데, 사고 건수 비율로 색깔을 포함해서 나타내고
신호등 위치를 나타낸 시각화_(원래는 회귀분석에서 유의한 변수인 교통안전표지판을 기준으로 시각화를 해야하는데 신호등이 유의한 변수라 착각하고 정리 중 알게 되어 고치지 못했습니다)



ii. 시각화

건물의 도로명 주소를 알 수 있는 '23.대전광역시_도로명주소(건물).geojson'을 신호등 분포와 함께 찍은 시각화



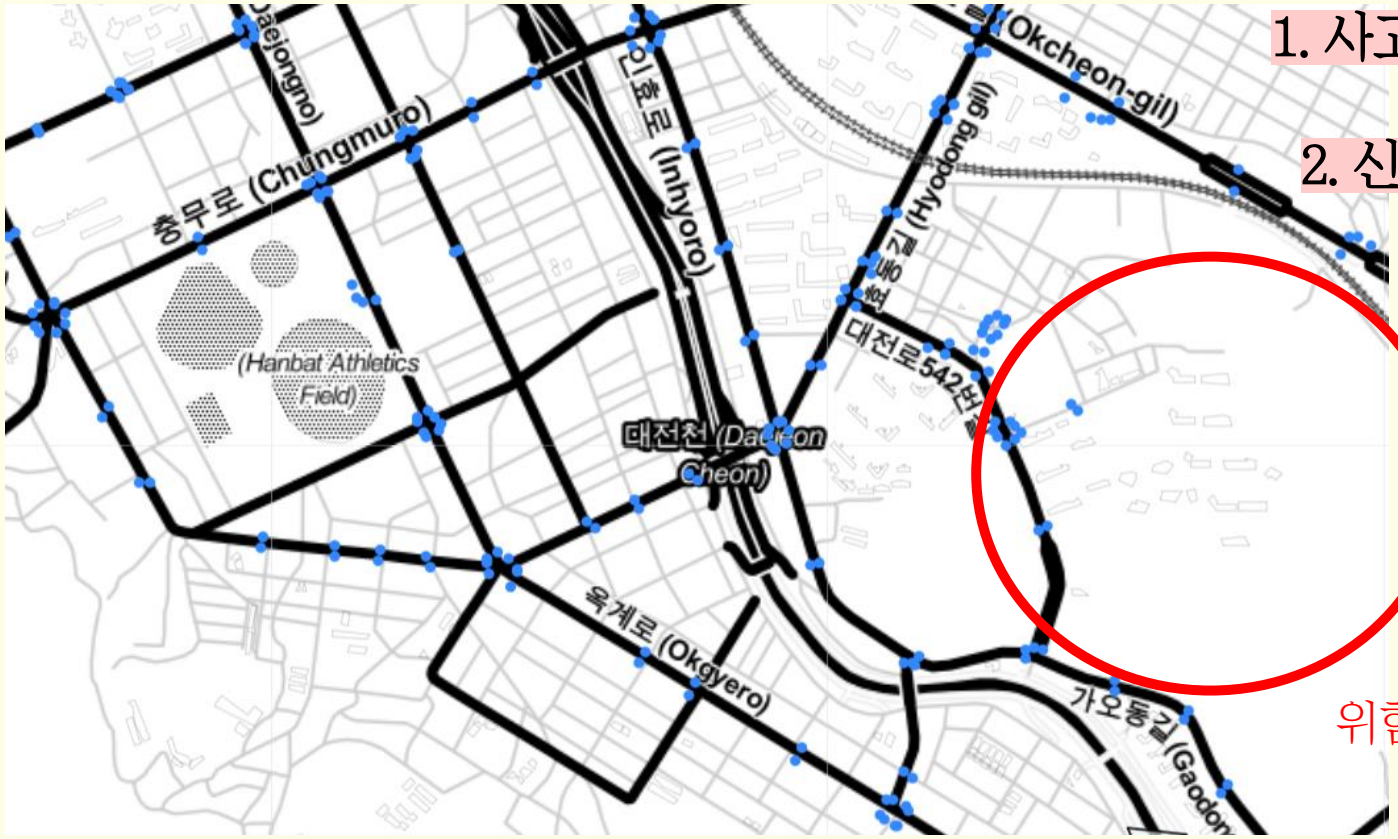
Ⅲ 결론

a. 결과해석 및 시각화
및 위험지역 도출

3. 결론

ii. 시각화해석 및 위험지역 도출

신호등이 유의한 변수로 착각하여 신호등이 없는 지역을 알아보고 그 지역의 건물을 파악하여 위험지역 100곳을 선정하였습니다.



최종적으로

1. 사고 건수가 많은 지역 중 신호 등 빈약 장소 선정
2. 신호등 개수가 적은 동 40개 중 세부 지역 선정

위험지역이라 판단

2

감사합니다

3

3