데이터 분석 최종결과보고서

I. 참가자 정보

제 목	환경적 요인과 교통사고 발생률의 상관관계 분석 및 보조 시스템	
팀 명	정민준	
성 명	정민준	
연락처	휴대폰	010-9391-0801
	E-mail	jmj284@gmail.com

II. 개요

ㅇ (분석/시각화) 목적

대전・세종・충남 지역 교통사고 데이터를 분석 후 시각화해서 데이터들 사이의 유의미한 관계를 찾아낸다. 찾아낸 관계와 데이터를 활용해서 경찰 행정 시스템을 보조하는 시스템을 제작한다. 한정된 경찰력을 효율적으로 배치할 수 있게 한다. 선제 대응 및 시대적 여건에 부응하는 맞춤형 치안 서비스 제공해서 사고 발생률을 낮춰서 지역 발전에 이바지한다.

ㅇ 배경 및 필요성

오늘날 대한민국의 교통안전 수준은 OECD 회원국 중 하위권에 속한다. 2021년의 분석 결과를 보면 10만 명당 교통사고 사망률은 5.9로 34개국 중 26위에 해당한다. 보행자 및 운전자로서의 경험 및 제공된 데이터의 분석 내용을 바탕으로 교통사고에 선제 대응을 할 수 있다. 단속 및 안전 시설물 설치 등의 행정적인 절차로 교통사고 발생률을 유의미하게 낮출 수 있다. 본 프로젝트에서는 단속 및 안전 시설물 설치 등에 동원되는 한정된 경찰력 및 자원을 효율적으로 배치하기 위한 보조 시스템을 제작하는 것을 목표로 한다. 결론적으로 이는 시간과 자원을 절약할 수 있다. 또한 교통사고 발생률도 유의미하게 감소할 수 있을 것으로 예상된다.

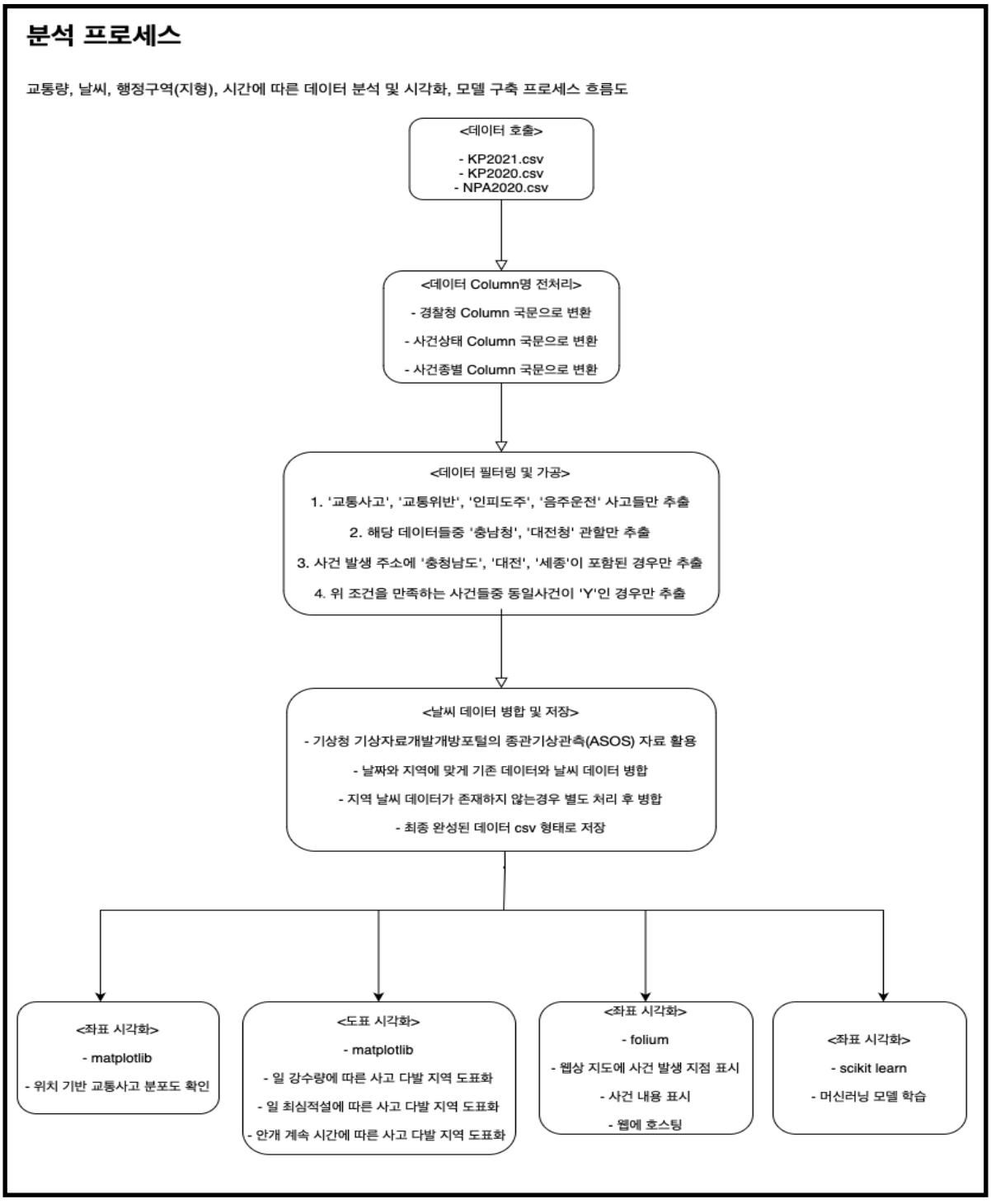
ㅇ 분석/시각화 결과 내용 요약

교통사고 발생과 유의미한 관계가 있는 요소들이 무엇인지 생각해 볼 필요가 있다. 운전자 입장에서 생각한 요소는 교통량, 날씨, 행정구역(지형), 시간 이렇게 네 가지를 꼽을 수 있다. 실제 인구가 많고 교통량이 많은 곳은 교통사고 발생률이 높았다. 또한 데이터 분석 결과 행정동으로 구분했다. 이때 날씨에 따른 사고 다발 행정동과 시간에 따른 사고 다발 행정동의 결과가 다양함을 알 수 있었다. 이것은 시간과 날씨에 따라서 경찰력을 투입해야 하는 곳이 달라짐을 의미한다. 이러한 데이터 분석 내용을 바탕으로 표와 지도를 기반으로 한 웹페이지를 제작한다. 또한 날씨와 시간에 따라 의사결정을 보조하는 머신러닝 모델을 제안해서 행정 시스템을 보조한다. 지역 지리 및 환경에 익숙한 지역 경찰관의 경험과 함께 데이터 분석 결과를 참고하면 효율적인 의사 결정에 도움이 될 것으로 예상된다.

III. 분석/시각화 결과 상세내용

0. 가설

교통사고에는 환경적인 요인이 많은 영향을 끼칠 것이라 가설을 정의한다. 교통사고의 요인에는 지형, 날씨, 시간 즉 외부 환경적인 요인이 더 크다는 것이다. 해당 가설 검증을 목표로 데이터 분석을 진행한다.



<분석 프로세스 흐름도>

1. <데이터 전처리>

효율적인 분석 및 경진대회 목적에 맞는 정확도 향상을 위해 데이터 분석에 불필요한 데이터를 삭제하는 전처리 과정을 거친다. KP2020.CSV, NPA2020.CSV, KP20201.CSV 파일에 모두 동일한 과정을 거친다.

- '교통사고', '교통위반', '인피도주', '음주운전' 사고들만 추출
- 해당 데이터들중 '충남청', '대전청' 관할만 추출
- 사건 발생 주소에 '충청남도', '대전', '세종'이 포함된 경우만 추출
- 위 조건을 만족하는 사건들중 동일사건이 'Y'인 경우만 추출
- 'HPPN_PNU_ADDR' 컬럼에서 행정동만 추출후 'SEARCH_ADDR' 컬럼을 새롭게 생성 후 전체 데이터 병합 후 저장(kp_result_complete.csv)

2. <날씨 데이터 추가 및 병합>

교통수단으로 인한 사고는 날씨 및 지형과 밀접한 관계가 있다고 판단했다. 따라서 기상청 기상자료개발개방 포털 (https://data.kma.go.kr)의 종관기상관측(ASOS) 자료를 활용한다.

2020 충남,대전,세종 날씨 데이터

	지점	지점명	일시	일강수량(mm)	일 최심적설(cm)	안개 계속시간(hr)
0	129	서산	2020-01-01	0.2	0.0	0.0
1	129	서산	2020-01-07	42.2	0.0	0.0
2	129	서산	2020-01-08	1.3	0.0	0.0
3	129	서산	2020-01-13	0.0	0.0	0.0
4	129	서산	2020-01-19	0.0	0.0	0.0
5	129	서산	2020-01-20	0.0	0.0	0.0
6	129	서산	2020-01-22	1.4	0.0	0.0
7	129	서산	2020-01-27	0.9	0.0	0.0
8	129	서산	2020-01-28	0.0	0.0	0.0
9	129	서산	2020-02-04	0.5	0.0	0.0
10	129	서산	2020-02-12	8.7	0.0	0.0

<충남, 대전, 세종 날씨 데이터>

1) 대전광역시, 세종특별자치시, 충청남도의 2020/01/01 ~ 2023/01/31 기간의 일간 '일강수량(mm)', '일 최심적설(cm)', '안개 계속시간(hr)' 데이터를 활용한다.

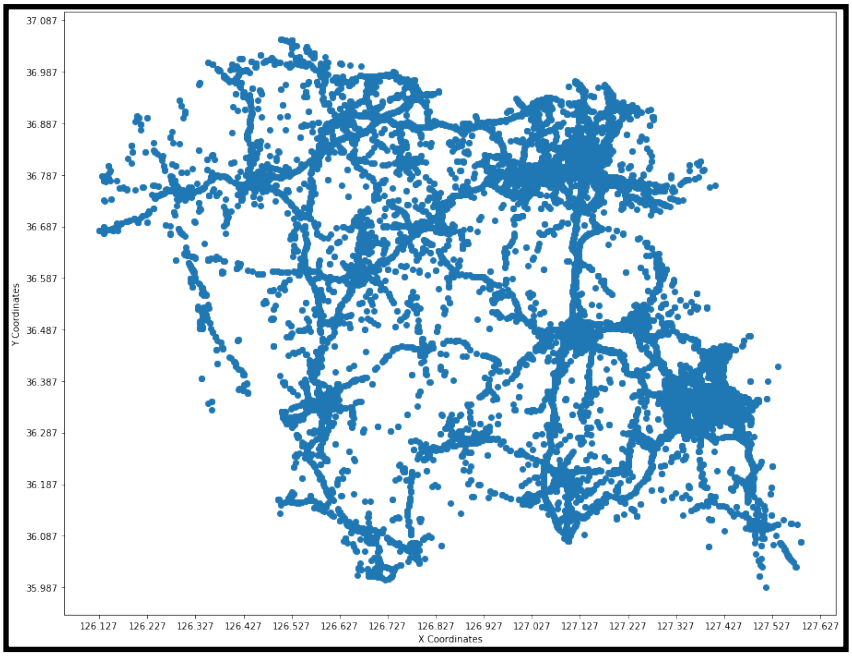
2) 날씨 데이터의 '일시'와 사건 데이터의 사건 발생 날짜('date')가 일치하고 날씨 데이터의 '지점명' 값이 사건발생 주소('HPPN_PNU_ADDR')에 포함되는 경우 해당 사건 데이터의 행에 '일강수량(mm)', '일 최심적설(cm)', '안개 계속시간(hr)' 열에 존재하는 데이터를 추가한다.

3-a) 해당 주소의 날씨 데이터가 존재하지 않는 경우 좌표 데이터를 바탕으로 해당 날짜의 가장 가까운 곳의 날씨를 Euclidean Distance 공식을 이용해서 가져온다. 해당 방법은 정확도가 높지만, 속도가 느리고 데이터의 크기가 작은 경우 유용하다. (KP2020.csv 데이터에 적용)

3-b) 해당 주소의 날씨 데이터가 존재하지 않는 경우 같은 날짜의 다른 지역의 날씨를 무작위로 가져온다. 무작위로 가져오는 이유는 순서대로 탐색하면 인덱스가 앞쪽에 있는 값만 계속 사용된다. 이는 데이터의 정확도를 감소시키고 모델링 과정에서 과적합(OverFitting)을 발생시킬 수 있다. 해당 방법은 정확도가 떨어지지만, 속도가 빠르고 데이터의 크기가 큰 경우 유용하다. (KP2021.csv 와 NPA2020.csv 데이터에 적용)

3. <좌표 시각화>

우선 교통사고 위치 데이터(HPPN_X, HPPN_Y)를 바탕으로 교통사고의 분포, 발생 요인 및 데이터 사이 관계의 이해를 돕기 위해 좌표를 시각화한다.



<HPPN_X, HPPN_Y 에 따른 교통사고 분포도>

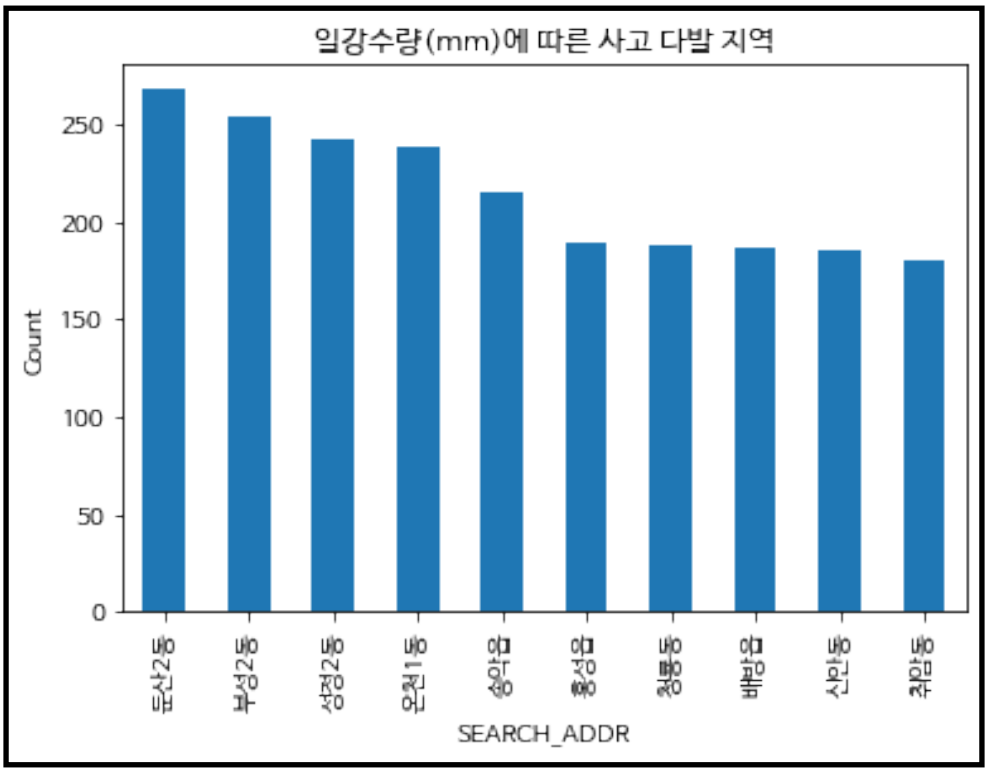
위의 분포도를 보면 특정 포인트에 데이터들이 모여있는 모습을 볼 수 있다. 분포도를 바탕으로 국내 지리 정보 및 인구수와 함께 비교해 본다. 사고가 많이 몰려 있는 곳을 몇 개 골라보면 서산시, 당진시, 대전광역시, 천안시로 볼 수 있다. 당진시와 서산시는 각각 2,400건 정도의 교통사고 통계가 있고 인구는 대략 17만 정도다. 천안시는 대략 9,300건 정도의 교통사고 통계를 갖고 대략 66

만의 인구를 갖는다. 대전광역시는 대략 2만 건의 교통사고 데이터를 갖고 인구수는 대략 145만 명을 갖는다. 인구 수당 교통사고 발생 비율은 각각 0.0140, 0.0140, 0.0140, 0.0138 이다. 해당 데이터는 모두 소수 다섯째 자리에서 반올림했다. 인구 밀도의 관계없이 거의 동일한 사고 재발률을 가진다. 따라서 환경적인 요인이 더 많은 영향을 끼친다고 판단 할 수 있다.

4. <도표 시각화>

위에 추가한 날씨 데이터(일 강수량, 일 최심적설, 안개 계속 시간)를 바탕으로 교통사고와의 상관관계를 matplotlib를 활용해서 도표로 분석한다.

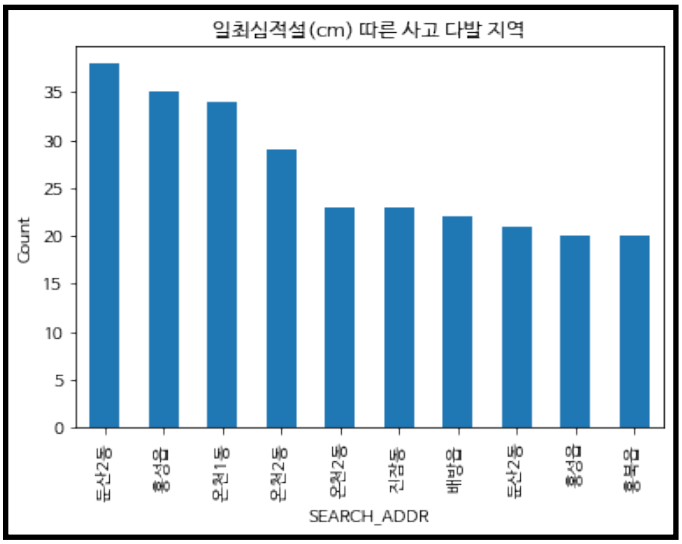
1. 일 강수량(mm)



<일 강수량에 따른 사고 다발 지역 도표>

일 강수량(mm)의 값이 0.0보다 큰 값을 가지는 데이터 중 행정동별로 사고 발생 횟수 순으로 정렬한 도표다. '둔산2동', '부성2동', '성정2동'이 비가 내리는 날 사고 발생확률이 가장 높은 세 곳임이 확인된다.

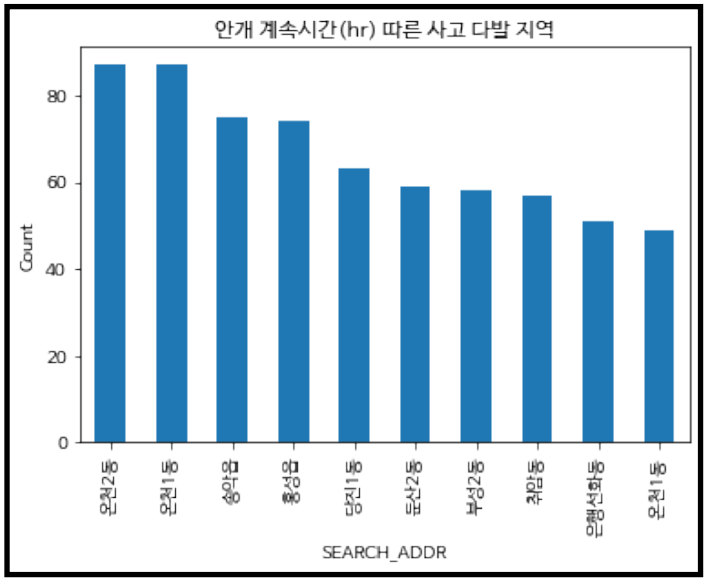
2. 일 최심적설(cm)



<일 최심적설에 따른 사고 다발 지역 도표>

일 최심적설(cm)의 값이 0.0보다 큰 값을 가지는 데이터 중 행정동별로 사고 발생 횟수 순으로 정렬한 도표다. '둔산2동', '홍성읍', '온천1동' 이 눈이 내리는 날 사고 발생확률이 가장 높은 세 곳임이 확인된다.

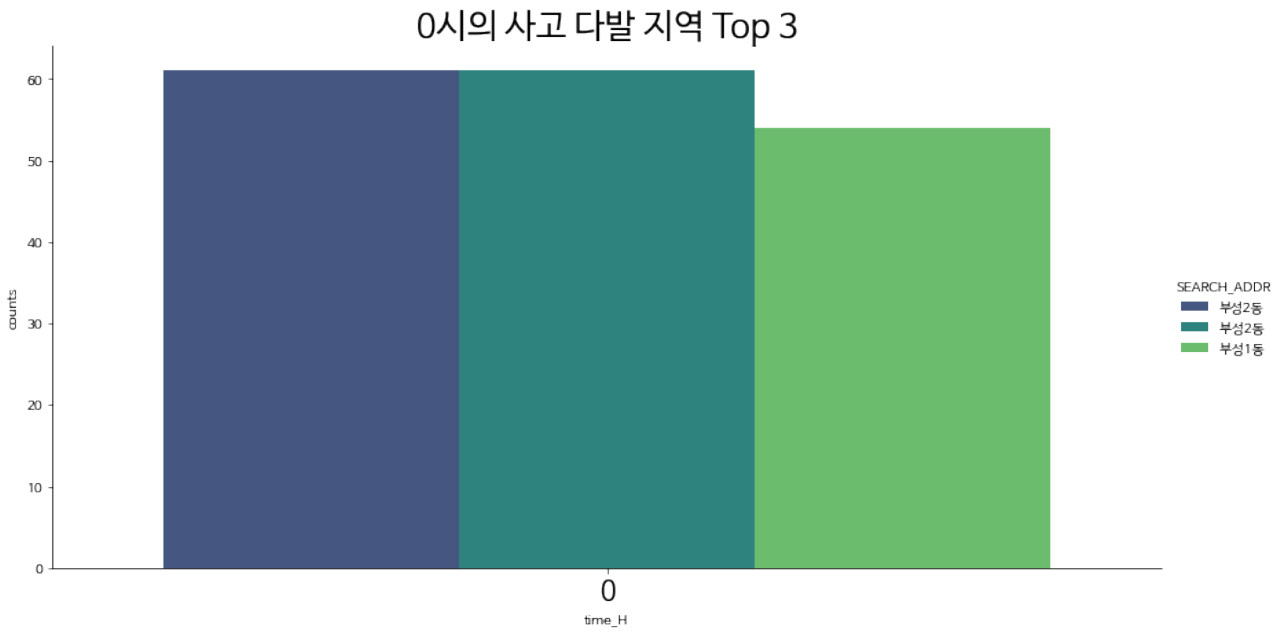
3. 안개 계속시간(hr)

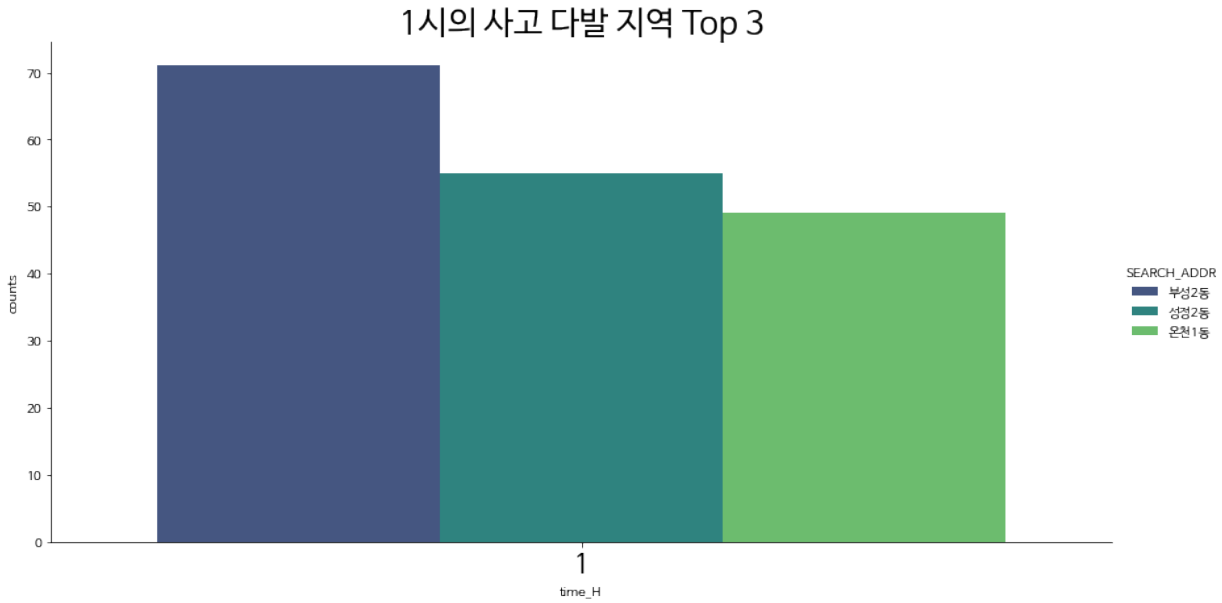


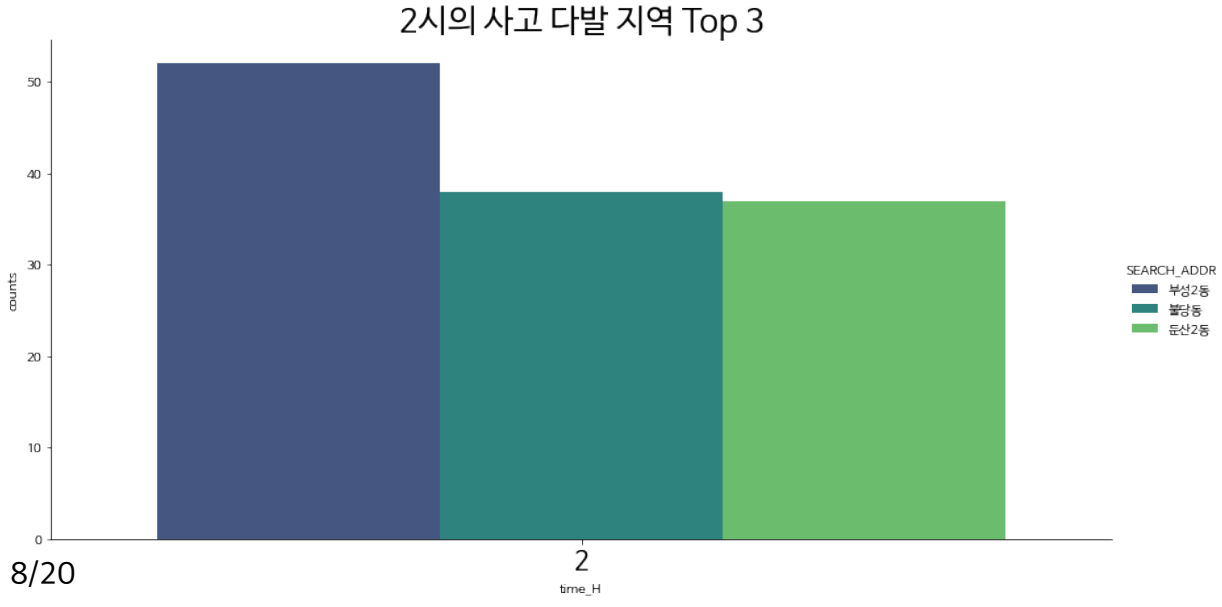
<안개 계속시간 에 따른 사고 다발 지역 도표>

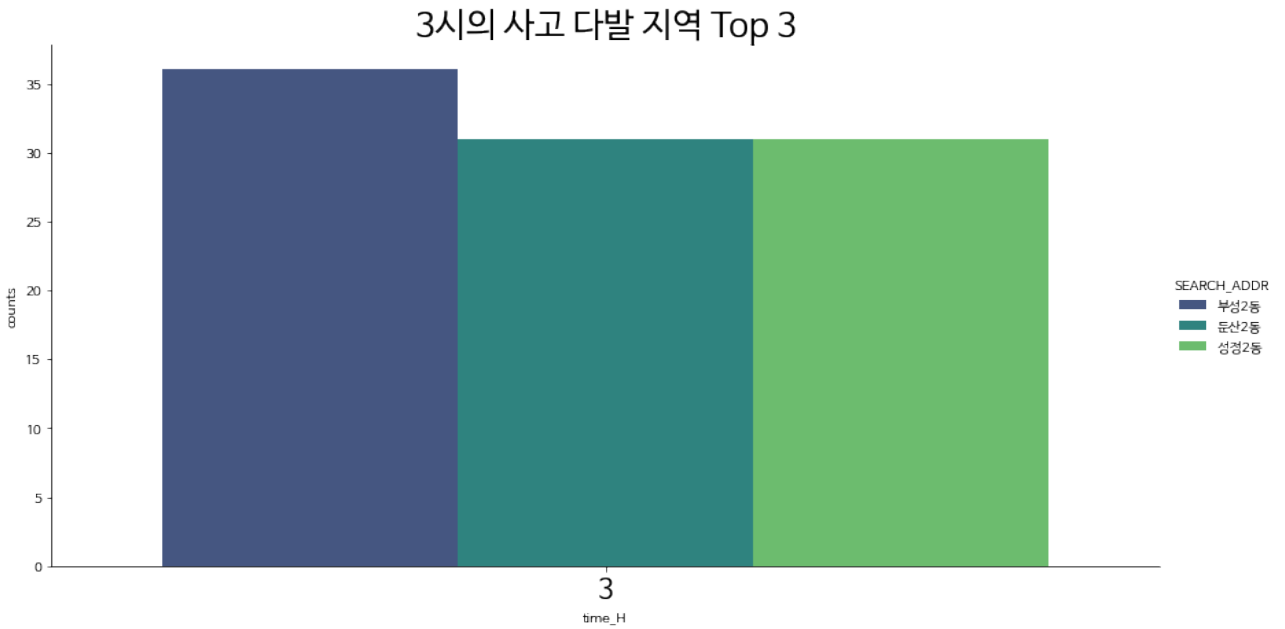
안개 계속시간(hr)의 값이 0.0보다 큰 값을 가지는 데이터 중 행정동별로 사고 발생 횟수 순으로 정렬한 도표다. ' 온천2동', '온천1동', '송악읍'이 안개가 낀 날 사고 발생확률이 가장 높은 세 곳임이 확인된다.

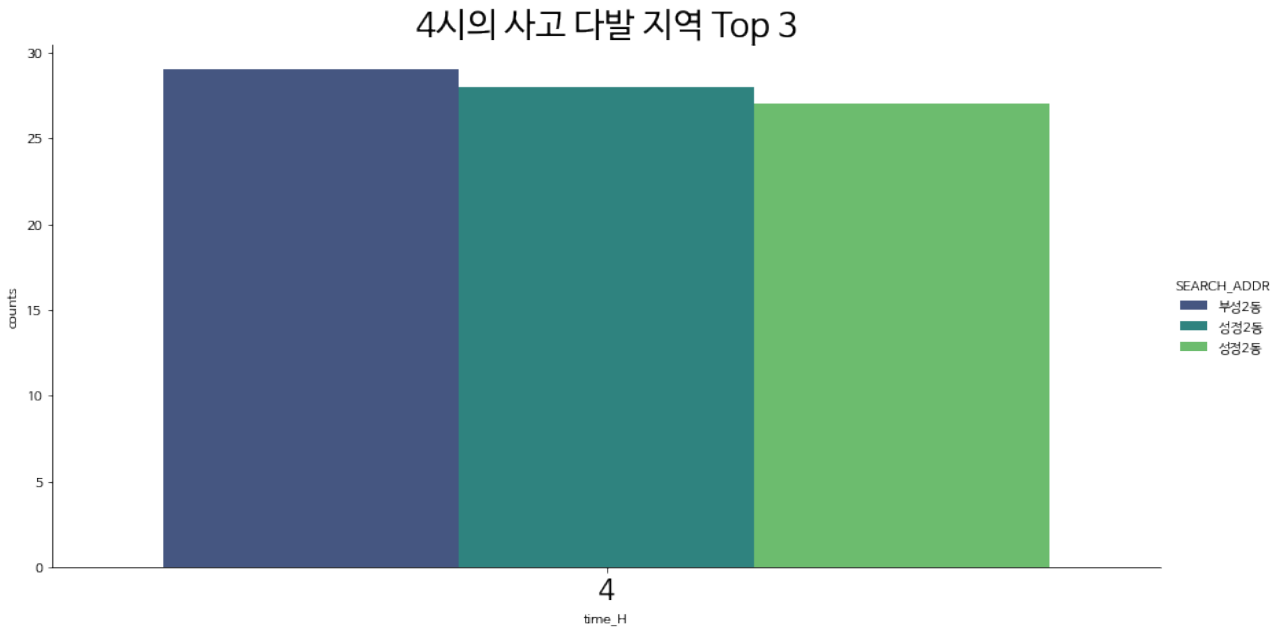
4. 시간에 따른 사고 다발 지역 차이

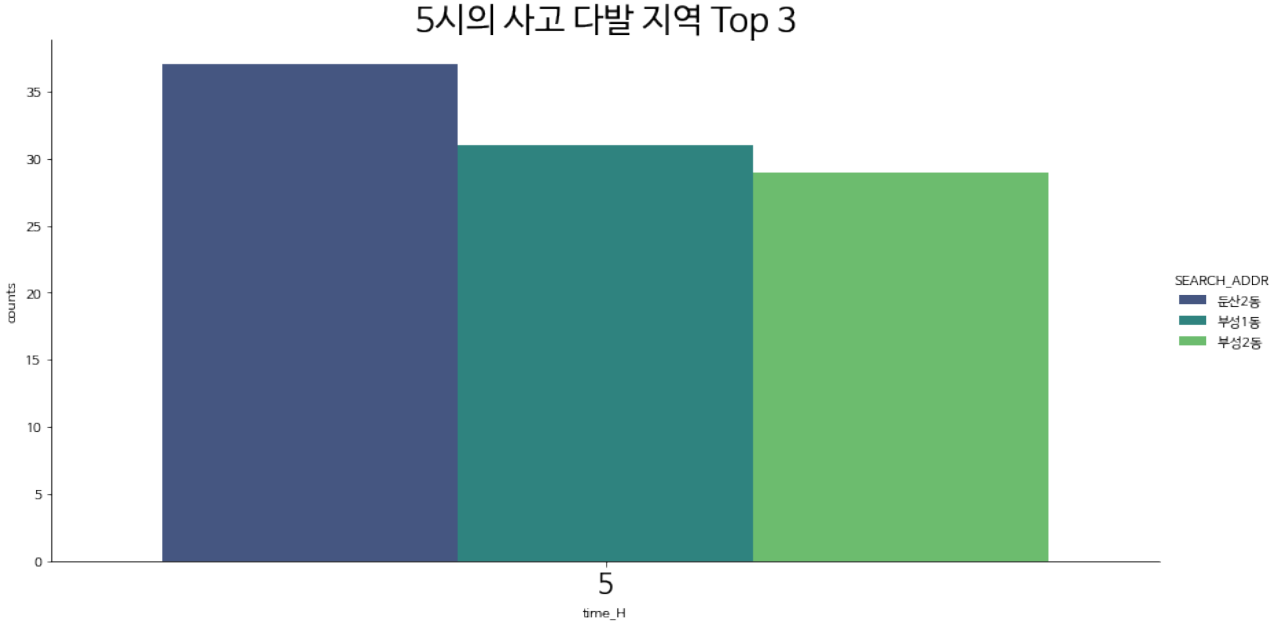


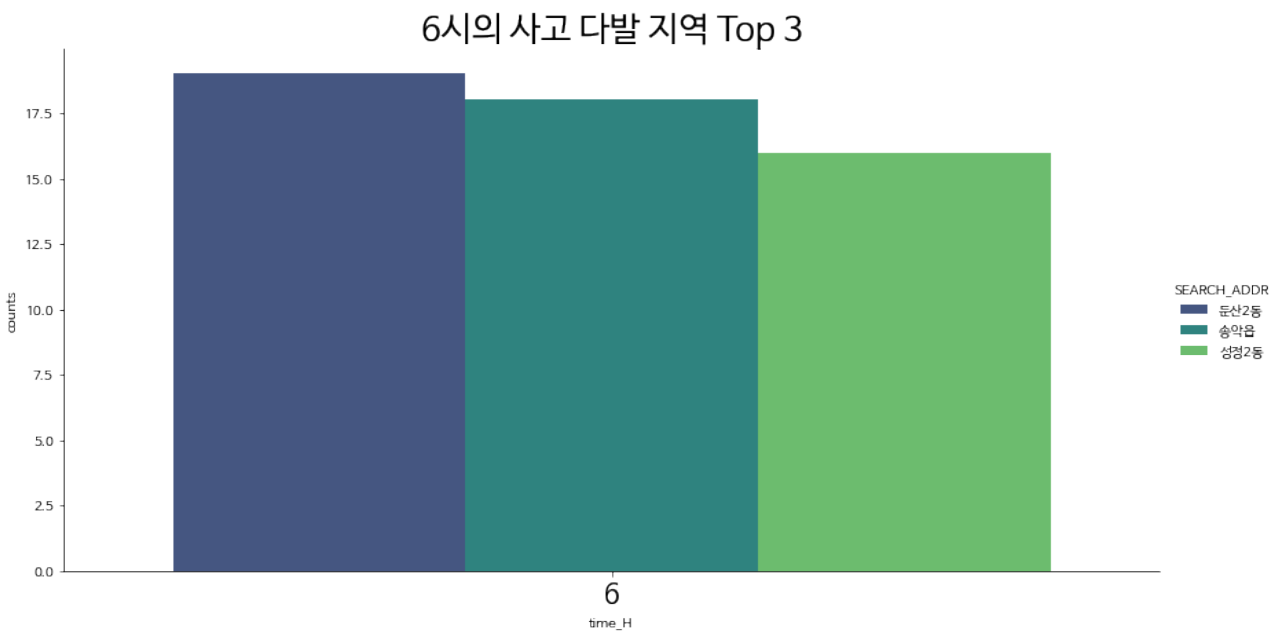


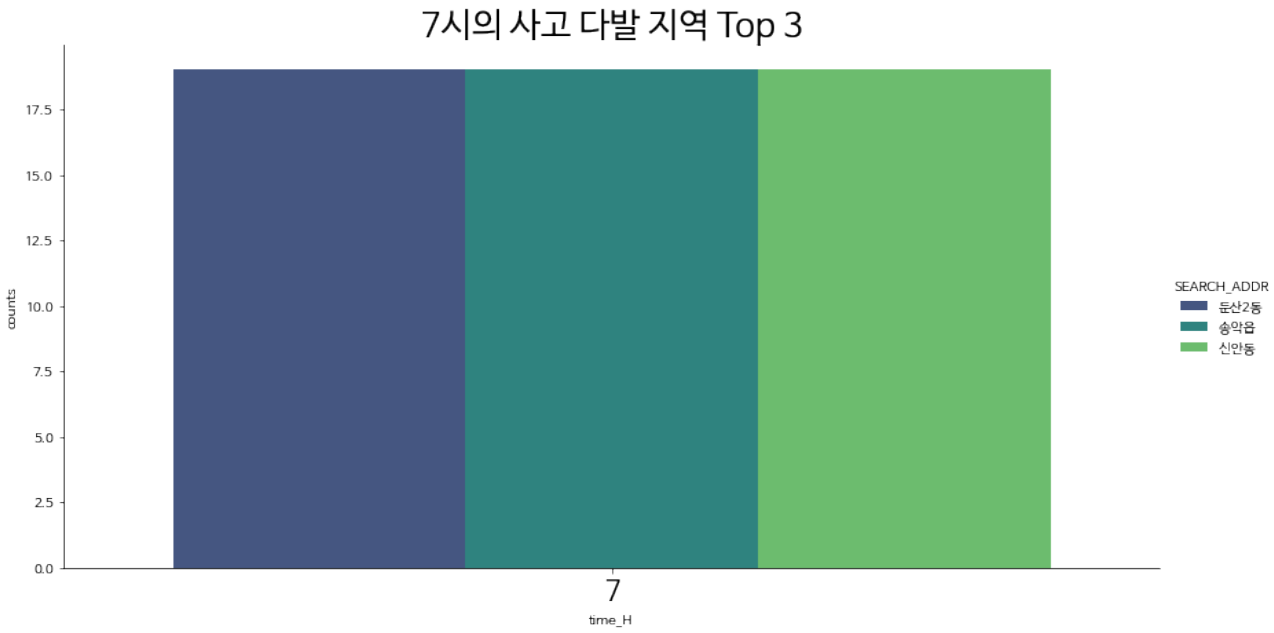


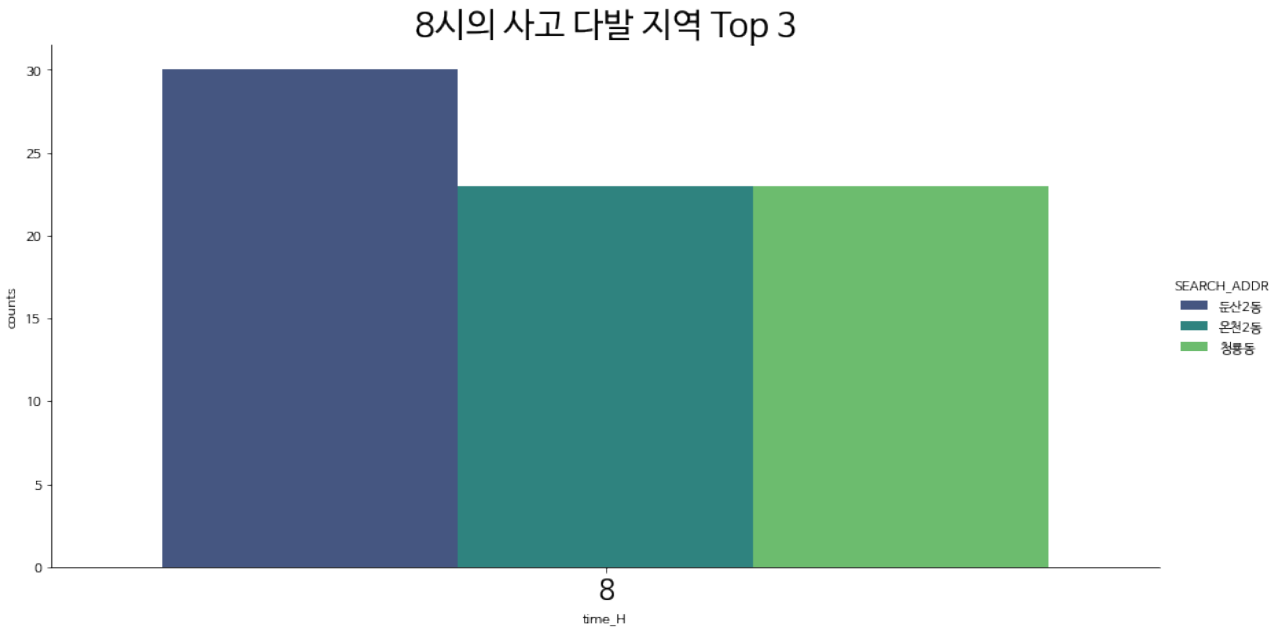

3시의 사고 다발 지역 Top 3
counts
time_H
3
SEARCH_ADDR
부성2동
둔산2동
성정2동

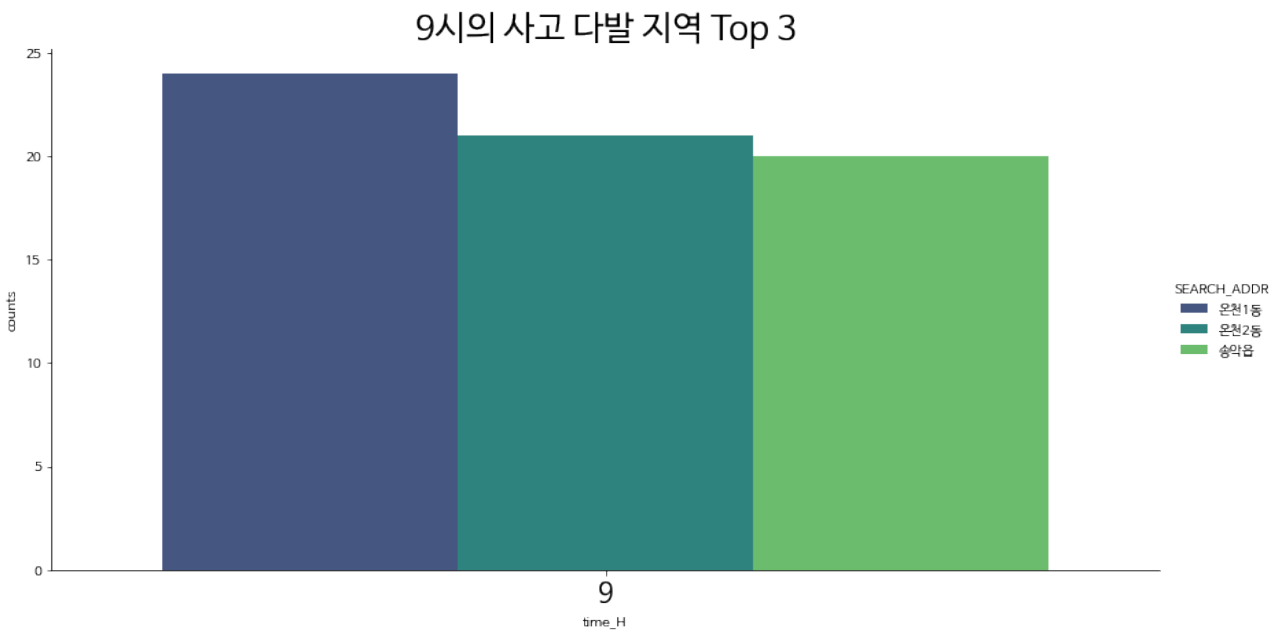

4시의 사고 다발 지역 Top 3
counts
time_H
4
SEARCH_ADDR
부성2동
성정2동
성정2동

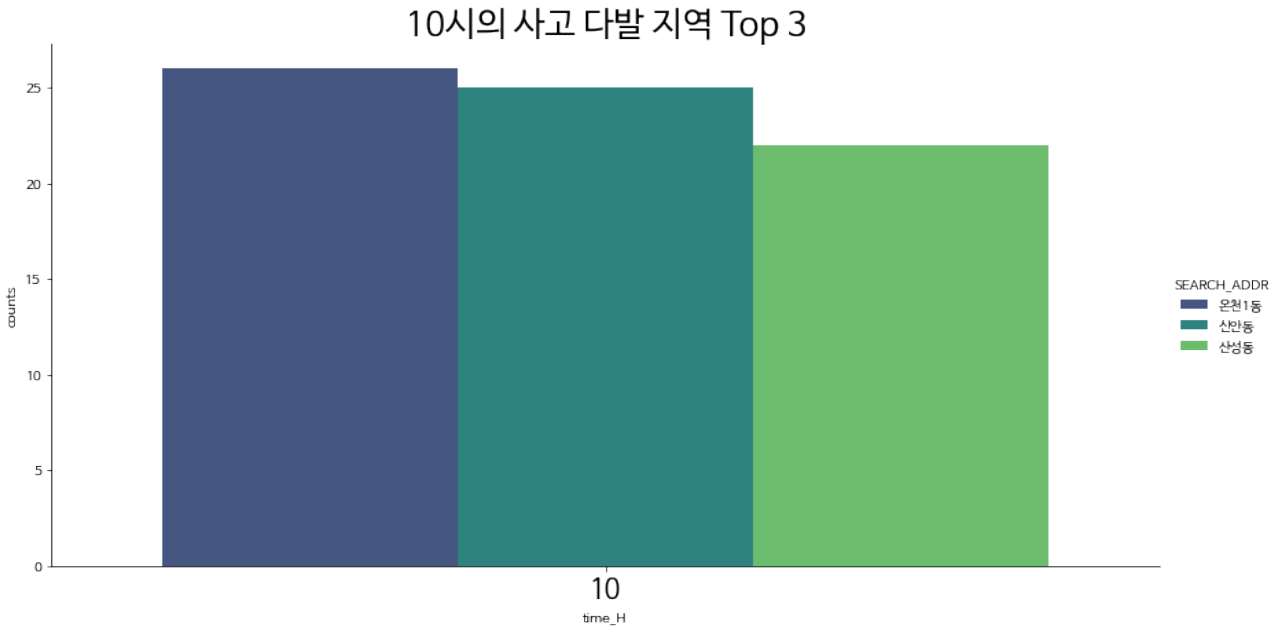

5시의 사고 다발 지역 Top 3
counts
time_H
5
SEARCH_ADDR
둔산2동
부성1동
부성2동

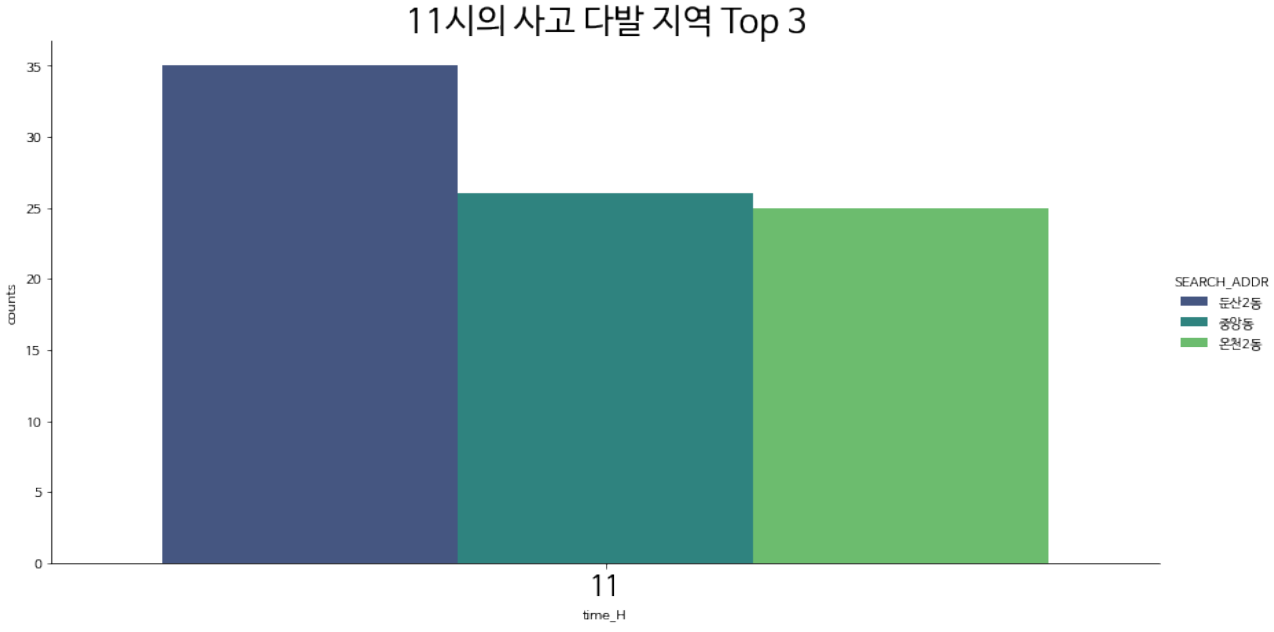

6시의 사고 다발 지역 Top 3
counts
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
6
time_H
SEARCH_ADDR
둔산2동
송악읍
성정2동

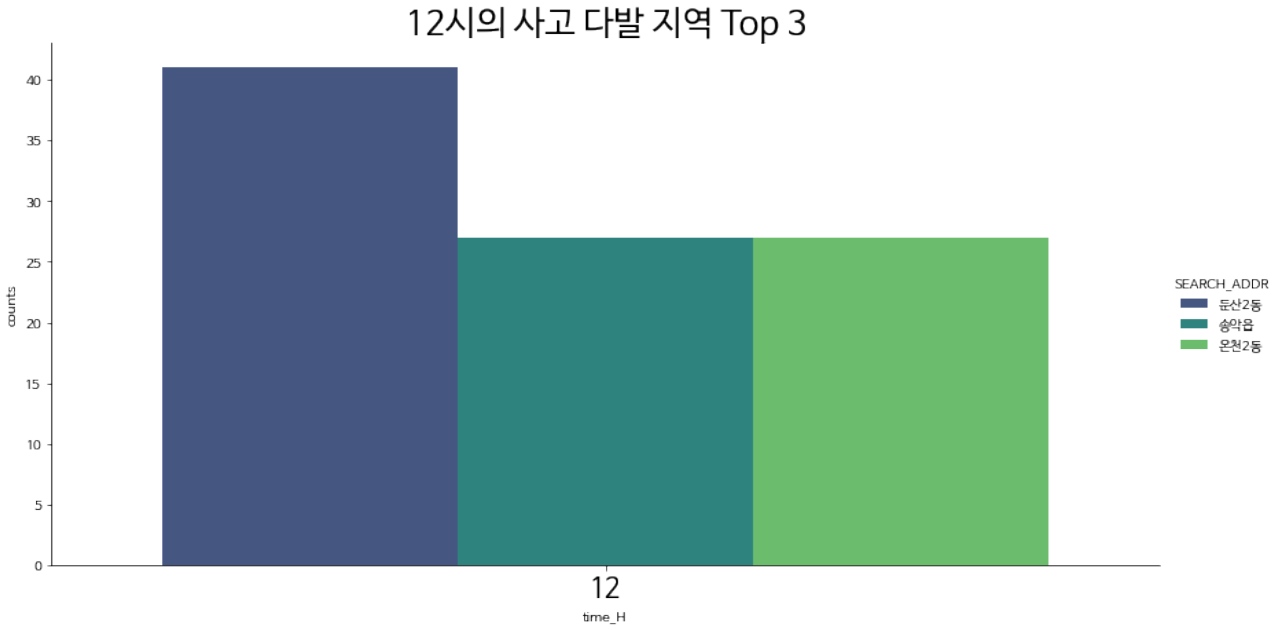

7시의 사고 다발 지역 Top 3
counts
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
7
time_H
SEARCH_ADDR
둔산2동
송악읍
신안동

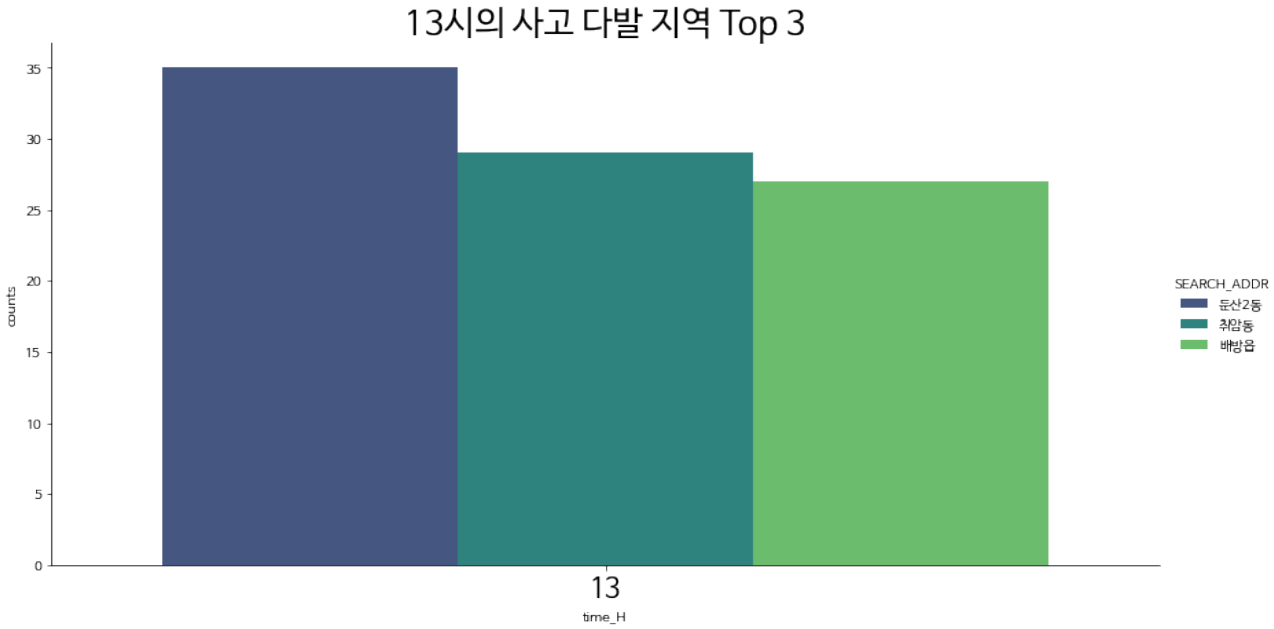

8시의 사고 다발 지역 Top 3
counts
30
25
20
15
10
5
0
8
time_H
SEARCH_ADDR
둔산2동
온천2동
청룡동

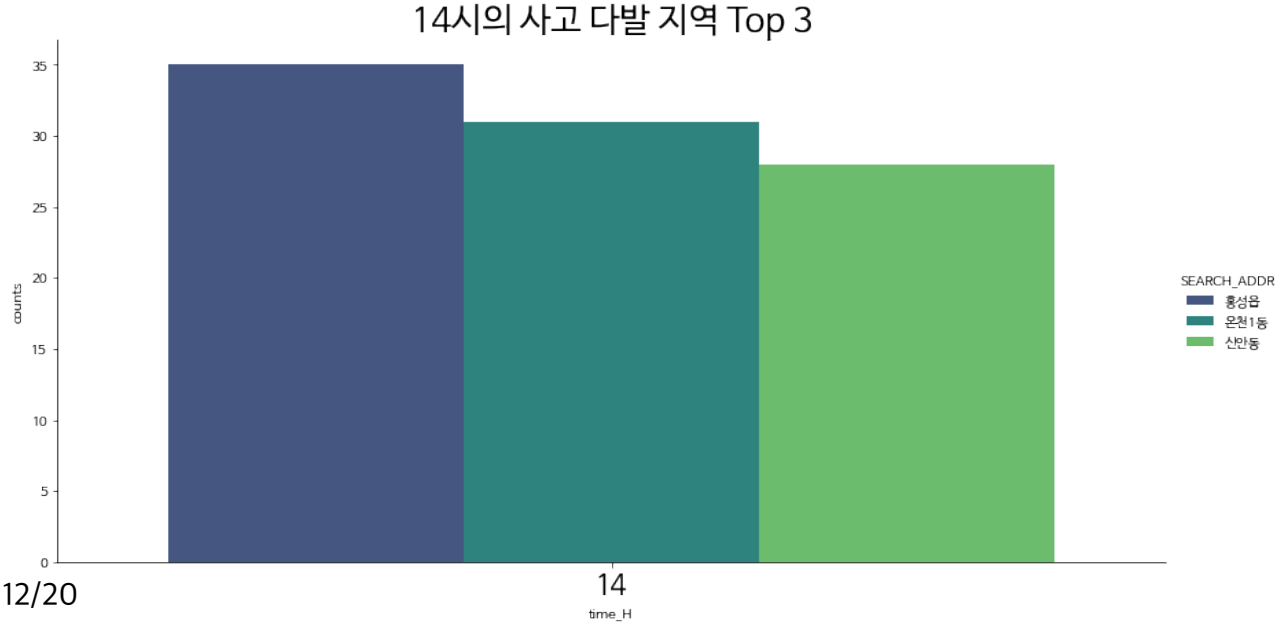

9시의 사고 다발 지역 Top 3
counts
time_H
9
SEARCH_ADDR
온천1동
온천2동
송악읍

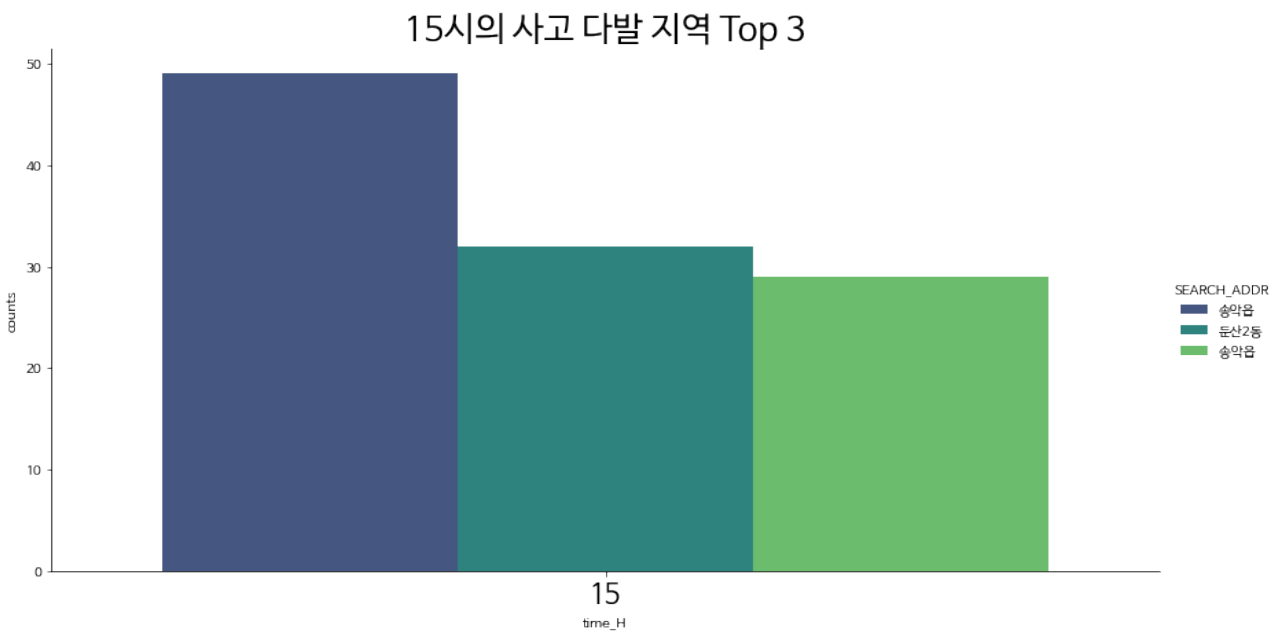

10시의 사고 다발 지역 Top 3
counts
time_H
10
SEARCH_ADDR
온천1동
신안동
신성동

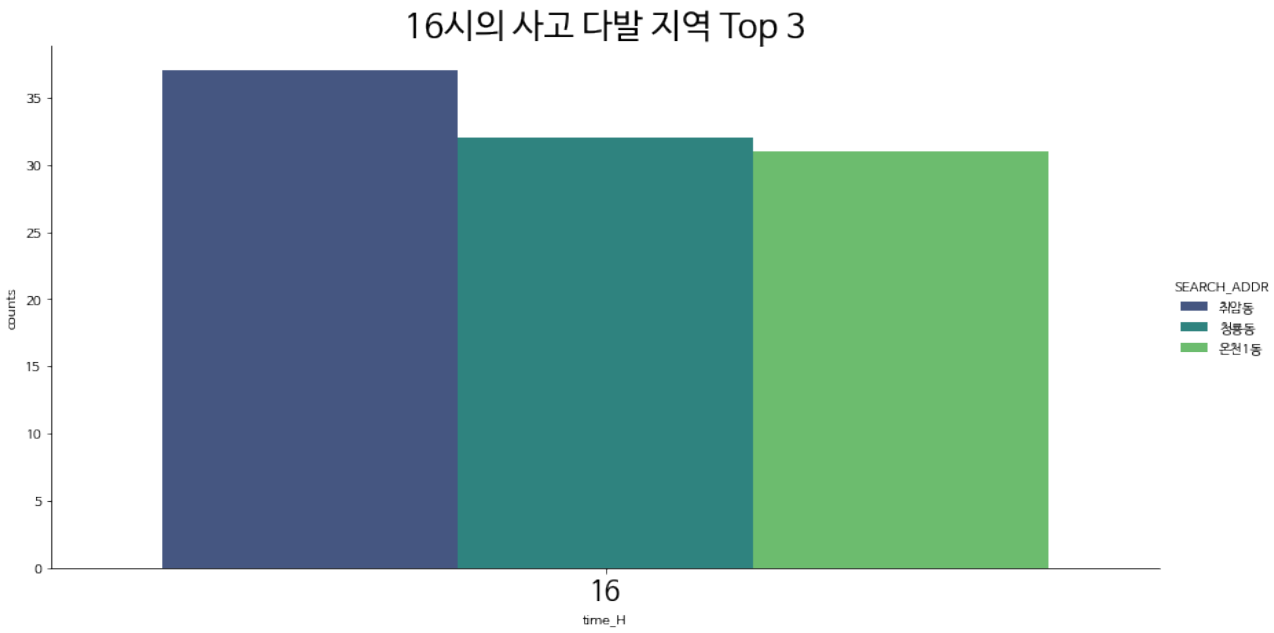

11시의 사고 다발 지역 Top 3
counts
time_H
11
SEARCH_ADDR
둔산2동
중앙동
온천2동

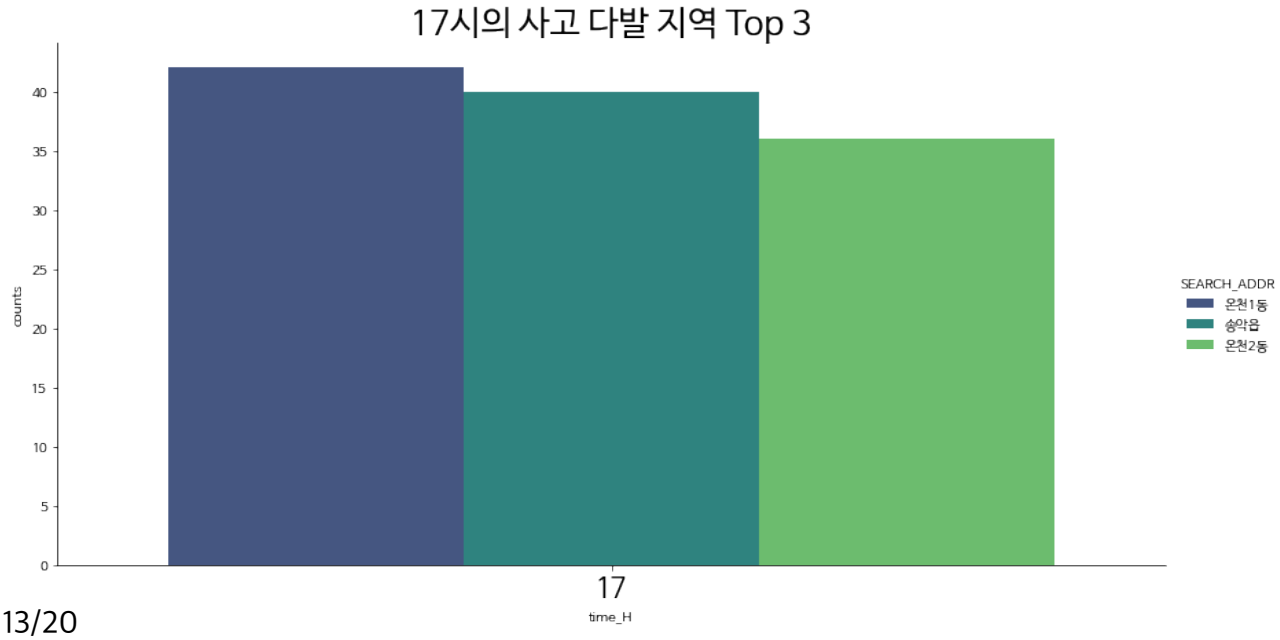

12시의 사고 다발 지역 Top 3
counts
0
5
10
15
20
25
30
35
40
12
time_H
SEARCH_ADDR
둔산2동
송약읍
온천2동

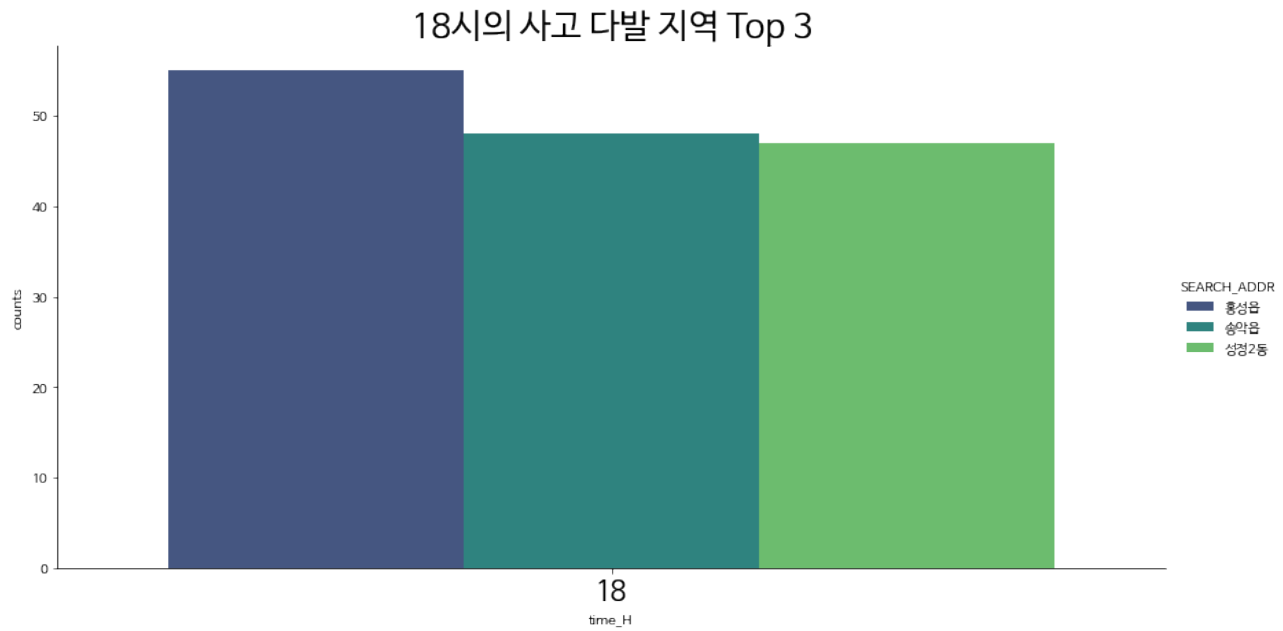

13시의 사고 다발 지역 Top 3
counts
0
5
10
15
20
25
30
35
13
time_H
SEARCH_ADDR
둔산2동
취암동
배방읍

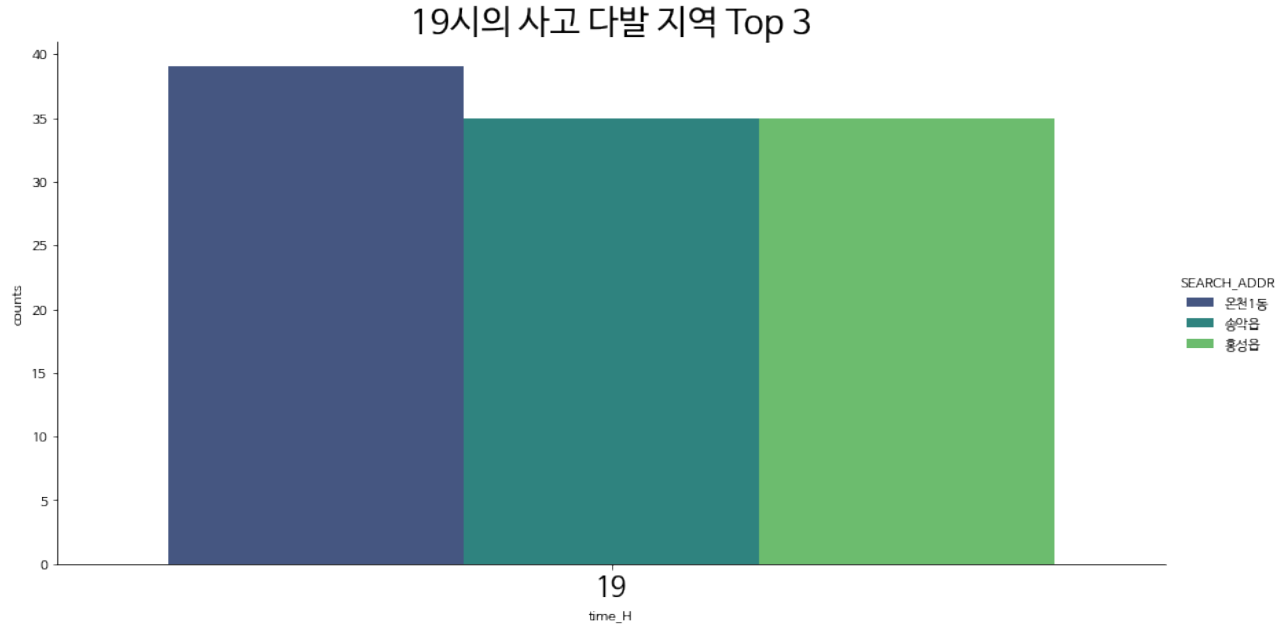

14시의 사고 다발 지역 Top 3
counts
0
5
10
15
20
25
30
35
14
time_H
SEARCH_ADDR
홍성읍
온천1동
신안동

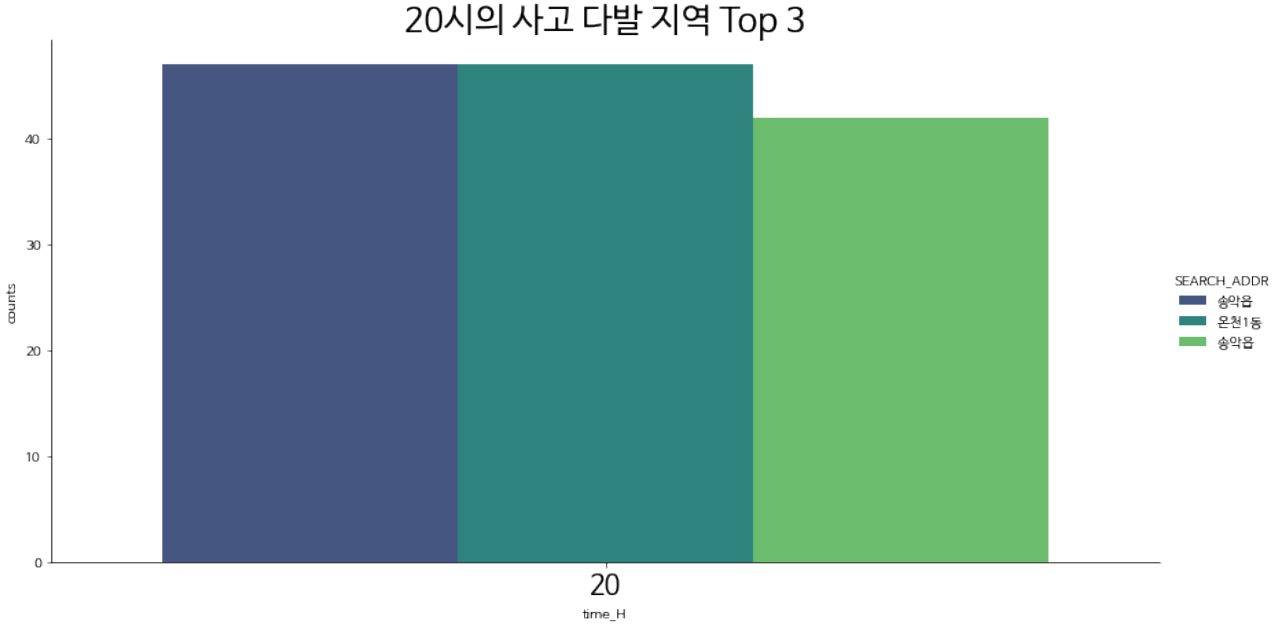

15시의 사고 다발 지역 Top 3
counts
50
40
30
20
10
0
15
time_H
SEARCH_ADDR
송악읍
둔산2동
송악읍

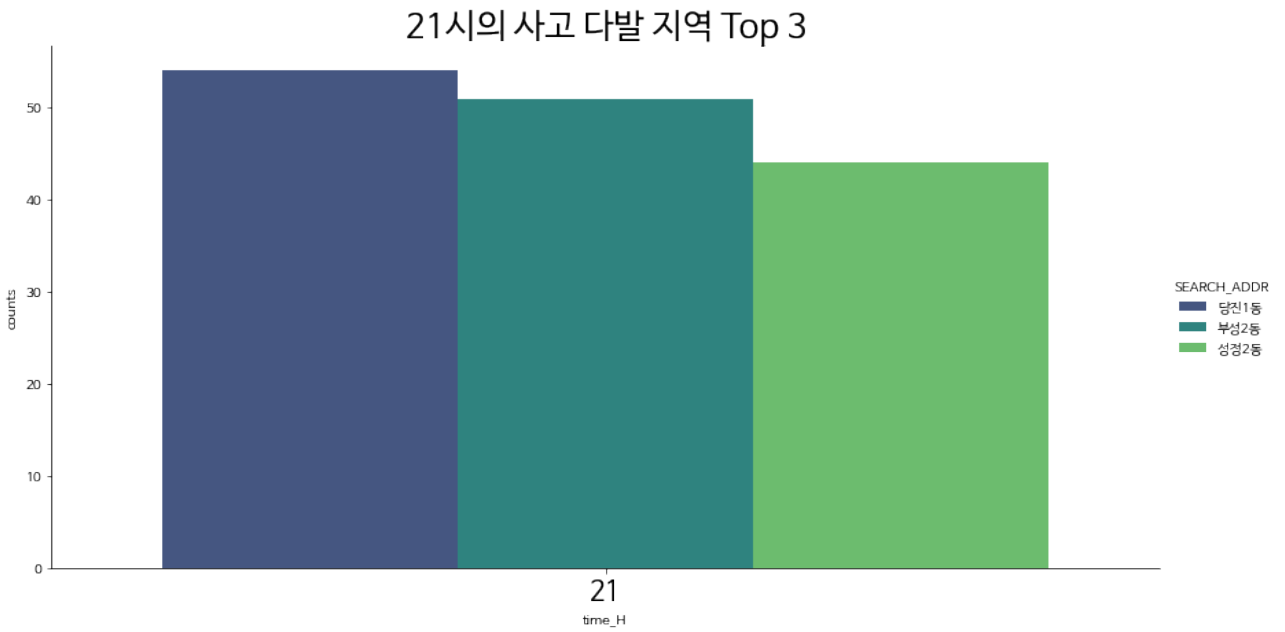

16시의 사고 다발 지역 Top 3
counts
35
30
25
20
15
10
5
0
16
time_H
SEARCH_ADDR
취암동
청룡동
온천1동

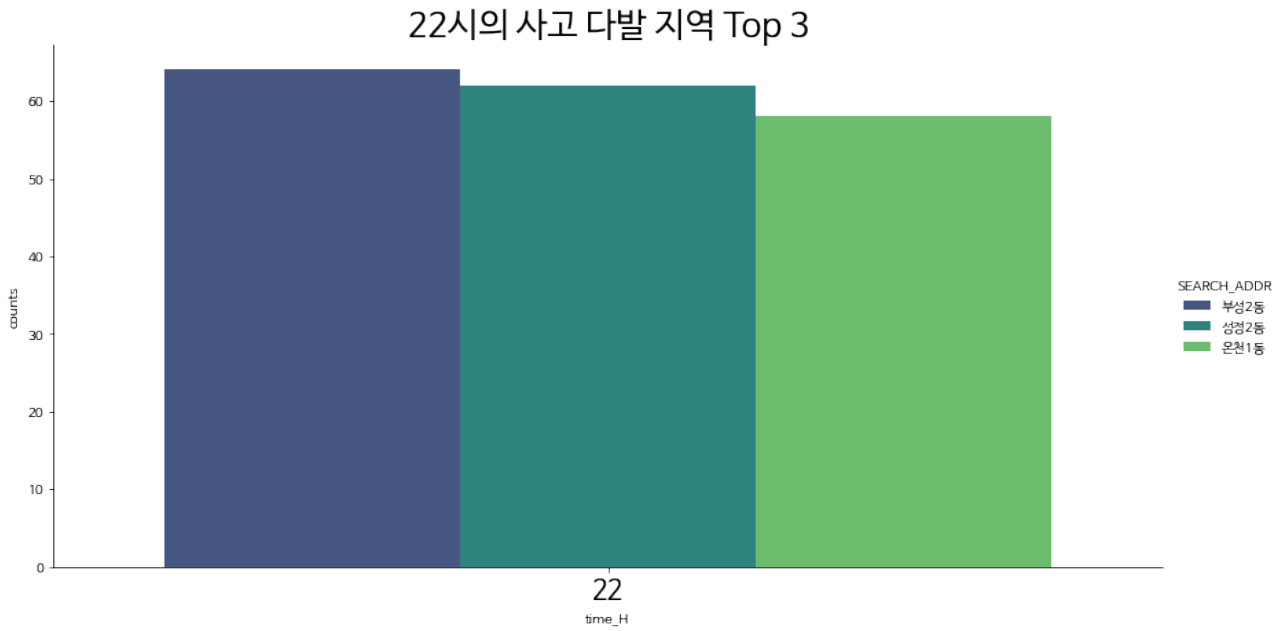

17시의 사고 다발 지역 Top 3
counts
40
35
30
25
20
15
10
5
0
17
time_H
SEARCH_ADDR
온천1동
송악읍
온천2동

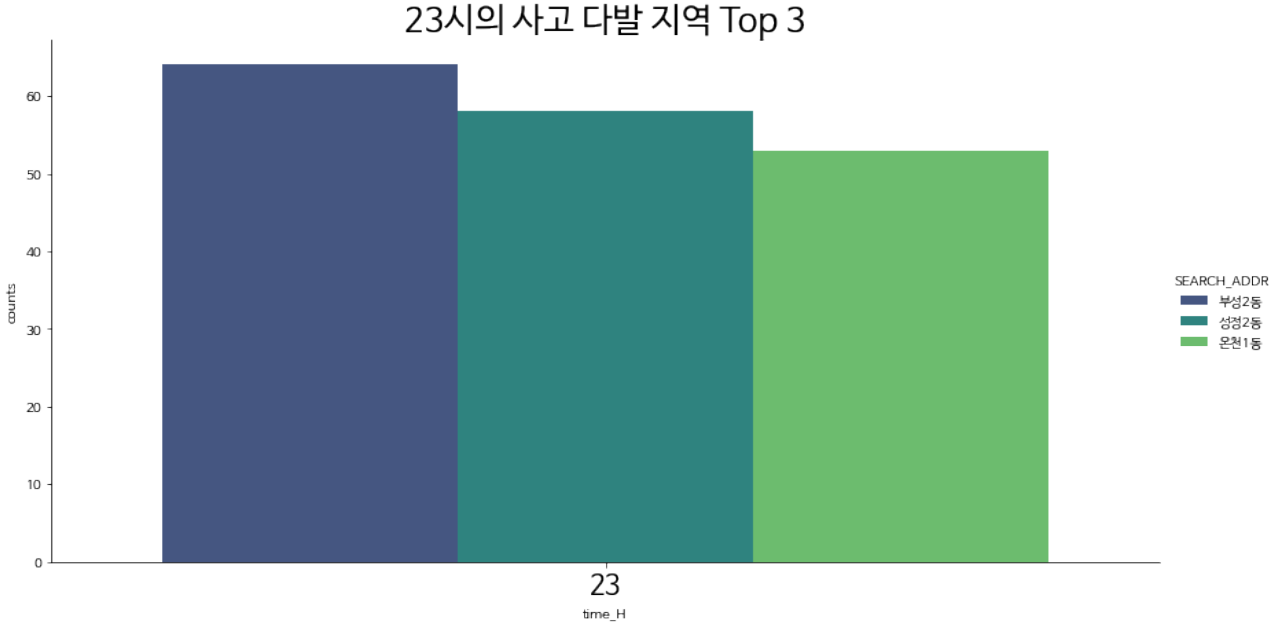

18시의 사고 다발 지역 Top 3
counts
50
40
30
20
10
0
18
time_H
SEARCH_ADDR
흥성읍
송악읍
성정2동


19시의 사고 다발 지역 Top 3
counts
40
35
30
25
20
15
10
5
0
19
time_H
SEARCH_ADDR
온천1동
송악읍
흥성읍


20시의 사고 다발 지역 Top 3
counts
40
30
20
10
0
20
time_H
SEARCH_ADDR
송악읍
온천1동
송악읍


21시의 사고 다발 지역 Top 3
counts
0
10
20
30
40
50
21
time_H
SEARCH_ADDR
당진1동
부성2동
성정2동


22시의 사고 다발 지역 Top 3
counts
0
10
20
30
40
50
60
22
time_H
SEARCH_ADDR
부성2동
성정2동
온천1동


23시의 사고 다발 지역 Top 3
counts
0
10
20
30
40
50
60
23
time_H
SEARCH_ADDR
부성2동
성정2동
온천1동

위의 도표를 보면 0시~23시 동안 시간별로 사고 횟수를 기준으로 사고 다발 지역을 상위 3개를 도표화한 것이다. 각 시간별로 사고 다발 지역이 상이해지는 결과를 볼 수 있다.
- 밤, 새벽 시간에는 부성 2동, 퇴근 시간에는 온천 1동, 출근 시간에는 둔산 2동과 온천 1동에서 사고가 자주 발생하는 것을 볼 수 있다.

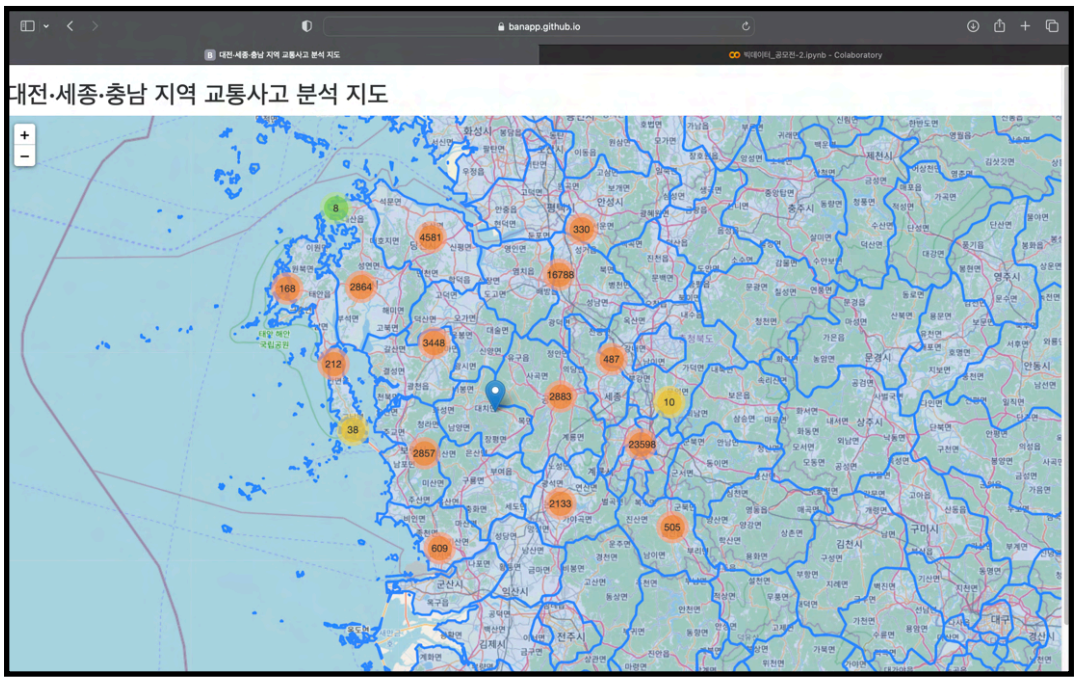
5. 도표 시각화의 결론 및 분석

위에서 소개한 일 강수량, 일 최심적설, 안개 계속시간을 통해서 얻은 결과를 표로 보면 각 데이터의 특성마다 결과가 유의미하게 다르다는 것이다. 즉, 비가 내릴 때 사고가 다발하는 지역, 눈이 내릴 때 사고가 다발하는 지역, 안개가 발생할 때 사고가 다발하는 지역이 다름을 의미한다. 또한 시간에 따라서 사고가 발생하는 지역이 유의미하게 달라진다. 따라서 분석된 해당 데이터를 바탕으로 지역과 시간을 고려해서 맞춤형 교통안전 대책 수립 방안을 제안한다.

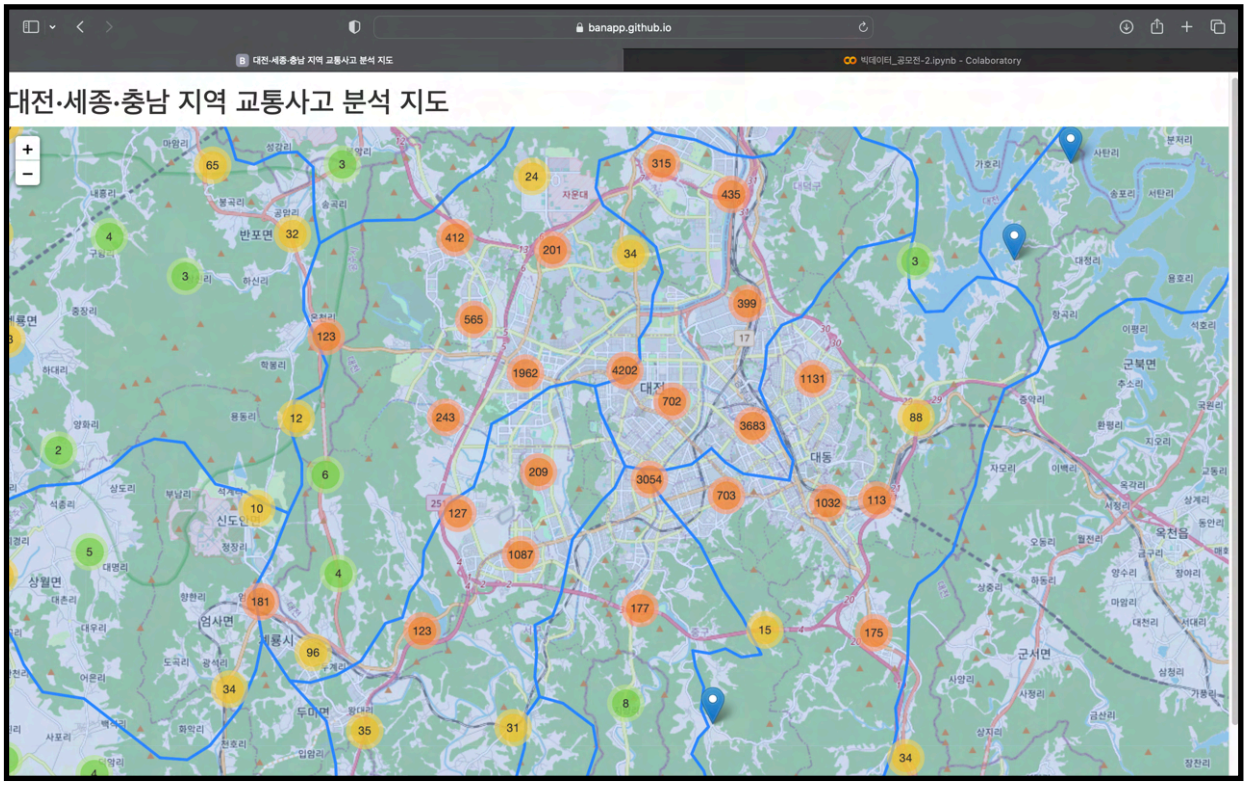
5. <지도 시각화>

파이썬의 지도 시각화 라이브러리인 'folium'을 활용해서 교통사고 정보를 실제 지도와 연동해서 시각화한다. 'HPPN_X','HPPN_Y'를 바탕으로 지도위에 핀(pin)을 보여준다. 각 핀에는 사건 종류('EVT_CL_CD'), 사건 발생 주소('HPPN_PNU_ADDR'), 사건 발생 날짜('RECV_CPLT_DM'), 일 강수량(mm), 일 최심적설(cm), 안개 계속시간(hr) 정보가 표시된다. 해당 서비스는 제작후 웹에 호스팅을 진행했다. 따라서 웹주소만 있다면 어디서든 접속할 수 있다.

주소: https://banapp.github.io/TAAM/



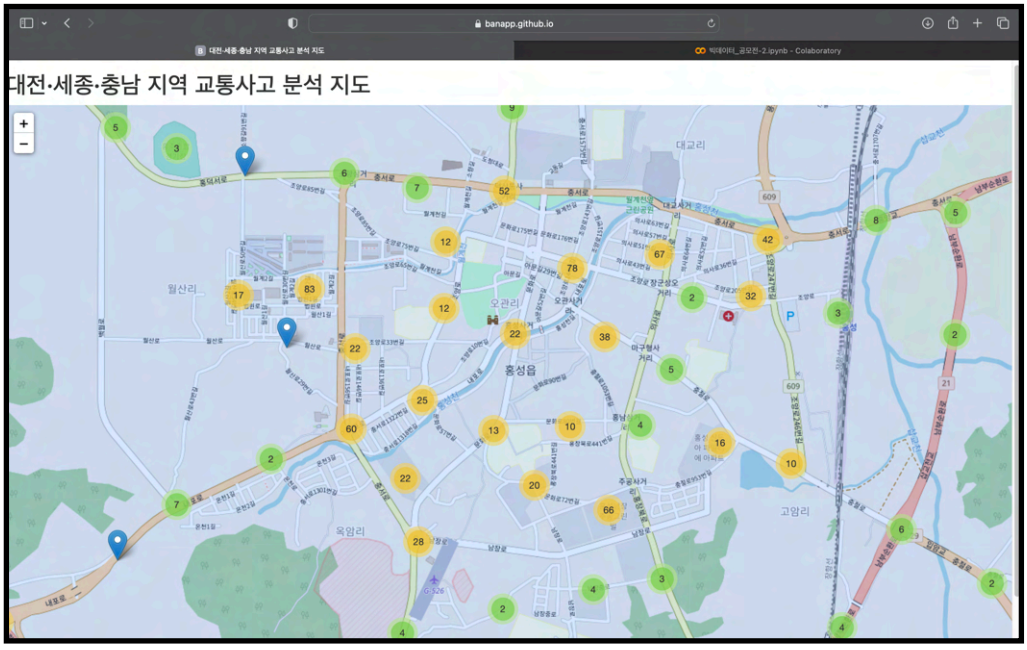
<웹 지도 오픈시 첫 화면>



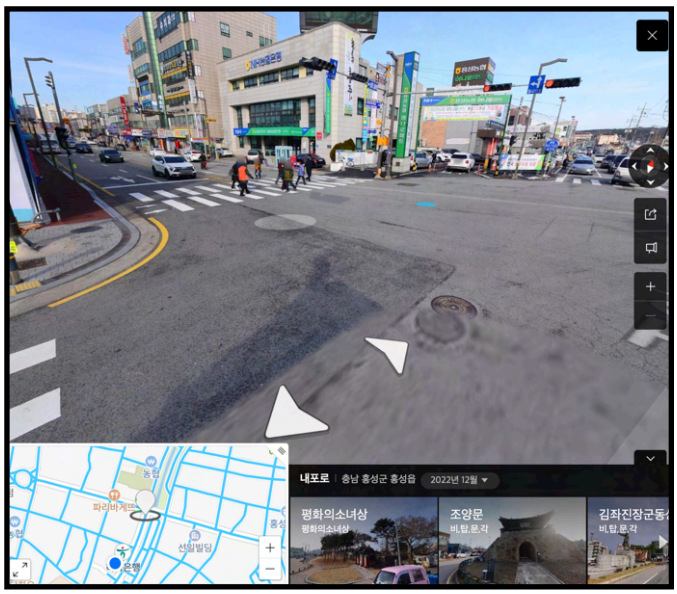
<지역별 교통사고 분포도 확인>

활용방안

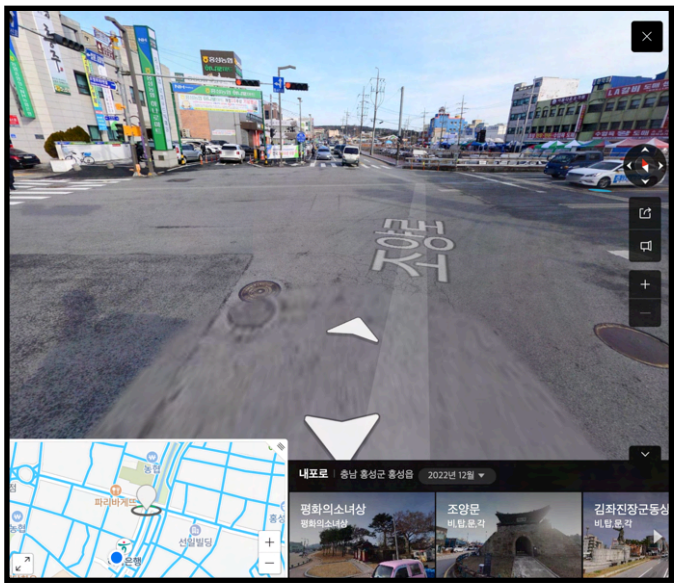
해당 지도 데이터를 활용하는 방안을 '홍성읍' 을 예시로 보인다.
앞의 '4. <도표 시각화>' 부분의 '2. 일 최심적설(cm)' 부분에서 눈이 내리는 경우 '홍성읍' 에서 사고 발생률이 높은 것을 확인했다. 따라서 웹 지도상의 홍성읍에서 사고가 자주 발생한 지점을 '네이버 거리뷰' 를 통해서 실제로 비교해 봤다.



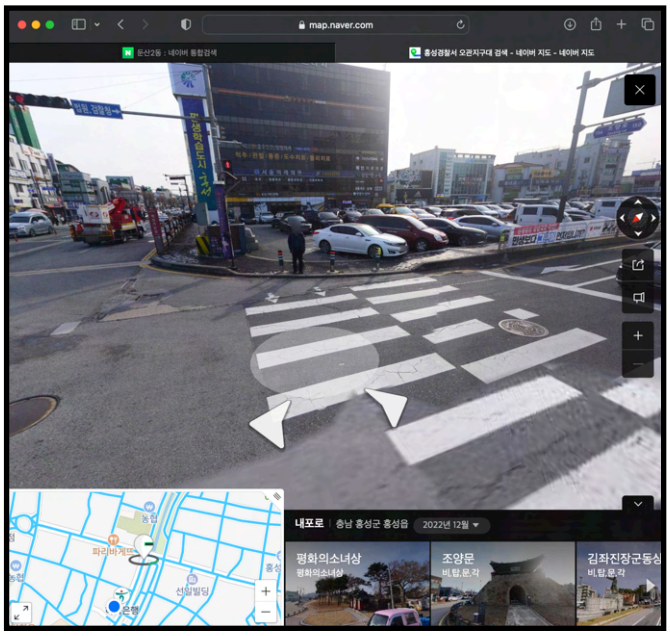
<웹 지도 상의 '홍성읍' 사고 분포도>



<홍성읍 조양로 거리뷰 - 2>



<홍성읍 조양로 거리뷰 - 1>



<홍성읍 조양로 거리뷰 - 2>

출처: https://map.naver.com/v5/search/홍성경찰서 오관지구대/place/17378670?c=16.35,0,0,0,dha&isCorrectAnswer=true&p=EXYKQLXRpj18GaX7XBQv2A,32.89,-24.02,80,Float

해당 도로는 번화가로 차량 및 보행자의 통행량이 많고 상가와 주차장 등이 많은 것에 비해 도로가 좁고 차가 많아서 사고의 위험이 매우 높아 보인다. 눈이 내리는 상황에서 도로 상태 개선에 사용될만한 대응 시설이나 환경이 부족해 보임에 따라서 상황에 따라 경찰력 투입으로 통제할 필요가 있다. 또한 식당 및 부대 시설이 많은 번화가이면서 교통량이 많기 때문에 교통 단속 및 음주 단속 지역으로도 적합하다.

o 결과 해석 및 시사점

정제되지 않은 빅데이터 속에서 목적과 필요에 맞게 데이터와 특징들을 분류하고 활용성을 높일 수 있도록 외부 데이터를 연계했다.

- 인구수 대비 사고율은 거의 동일하다. 따라서 외부 요인이 교통사고에 더 많은 영향을 끼친다.
- 비, 눈, 안개에 따른 날씨에 따라서 교통사고가 다발하는 지역의 순위가 유의미하게 바뀐 것을 확인할 수 있다.
- 날씨뿐만 아니라 시간에 따라서도 교통사고가 다발하는 지역의 순위가 유의미하게 바뀐 것을 확인할 수 있다.
- 따라서 지역과 시간에 따라서 단속 및 안전 설치물을 설치할 필요가 있다. 지역 맞춤형 정책이 필요하다.
- 해당 정책들을 보조하기 위해서 위에서 제작한 웹 기반의 지도와 머신러닝 모델 도입을 제안한다.

o 기대효과

데이터 분석 결과를 참고하고 이를 바탕으로 보조 도구를 제작해서 지역 경찰관의 역량과 더해지면 교통사고 발생률 감소에 유의미한 영향이 있을 것이다.

- 해당 데이터 분석을 통해서 교통사고를 유발하는 요인과 상관관계에 대해서 분석했다.
- 해당 분석 내용을 바탕으로 새로운 정책 및 계획 수립에 참고할 수 있다.
- 날씨, 시간에 따라서 교통사고 발생률이 달라질 수 있고 이를 참고해서 교통 단속 및 시설물 설치를 하면 경찰력을 효과적으로 운용해서 사고율을 낮출 수 있다.
- 웹 기반의 지도를 참고해서 실제 교통 다발 지역을 파악해서 해당 지역을 집중적으로 단속한다.
- 웹 기반의 지도는 웹에 존재 하므로 어디서든 인터넷이 되는 장소와 디바이스만 있으면 접속해서 사용할 수 있는 장점이 있다.
- 머신러닝 기반의 의사 결정 보조 인공지능 모델은 단속 및 교통 시설물 설치 지역을 선택하는 데 도움을 준다.
- 해당 도구들의 보조와 함께 해당 지역의 상황과 지리를 잘 알고 있는 현장 경찰관들의 역량이 더해지면 큰 시너지 효과가 생길 것으로 기대된다.

IV. 기타

ㅇ 건의 사항

데이터 분석 및 보고서 제작 과정에서 날씨 데이터(일 강수량, 일 최심적설, 안개 계속시간)를 바탕으로 사고 발생 가능성이 있는 머신러닝 모델을 제작해 봤다. X, Y좌표를 예측하는 회귀(Regression)모델을 제작하려 했으나 성능이 좋지 못해 분류 모델로 제작했다. Gradinet Boosting 기반의 트리 계열 모델로 제작했다. 그러나 날씨의 데이터가 해당 지역의 정보와 완전히 일치하지 않고 데이터의 결측치를 Euclidean Distance 같은 기초적인 방법으로 결측 데이터를 채웠기 때문에 정확도 부분에서 보장이 힘들다. 또한 참가자 개인의 컴퓨팅 자원으로 해당 규모의 모델을 제작 및 배포 하는 부분에서 한계가 명확했다.

날씨 데이터를 지자체에서 제공받고 충분한 분량의 사고 데이터, 시간, 날씨 데이터, 행정동 정보를 입력받아서 회귀 모델로 제작하면 출력값으로 해당 행정동 안의 사고 발생 가능성이 높은 좌표를 출력하는 모델을 만드는 것도 비용면과 실용면에서 따졌을때 가능성이 있다고 생각된다.

ㅇ 활용 데이터 및 참고 문헌 출처 (필수)

1. 대한민국 기상청 (2023). ASOS 실시간 기상 관측자료 [온라인]. 제공: data.kma.go.kr. [접속일: 2023년 2월 13일]. 웹사이트: https://data.kma.go.kr/data/grnd/selectAsosRltmList.do.

2. Naver Corp. (2023). 홍성읍 조양로 [온라인 지도]. 제공: map.naver.com. [접속일: 2023년 2월 13일]. 웹사이트: https://map.naver.com/v5/search/홍성읍%20조양로?c=15,0,0,0,dh&p=dr9AGYsnCEjSdb3aefQPPw,-87.41,7.27,80,Float.