장마 강수량과 호우피해 데이터 분석

목차

1. 서론

• 분석 계기

최근 기후변화로 인한 이상기온 현상으로 폭염과 폭우와 같은 자연재해가 빈번하게 발생하며 큰 피해를 초래하고 있습니다. 특히 호우 피해와 관련된 뉴스를 보며, 피해자들이 겪는 어려움을 접할 때마다 안타까운 마음이 들었습니다. 자연재해는 예측하기 어렵고, 피해를 입은 후 다시 일상으로 돌아가는 데는 막대한 비용과 오랜 시간이 소요됩니다. 이 때문에 복구보다는 예방이 무엇보다 중요하다는 생각이 들었습니다.

이번 분석은 장마 기간의 강수량과 호우 피해의 위험 요인들을 심층적으로 분석하여, 어떤 요인들이 피해 규모를 증가시키는지 파악하고자 진행되었습니다. 이를 통해 호우 피해를 예방하는 데 도움이 되기를 바라는 마음으로 데이터를 분석하게 되었습니다.

1. 서론

- 개요

본 프로젝트는 장마 기간의 강수량과 수해 피해 데이터를 분석하여 피해의 원인과 영향을 파악하고, 이를 바탕으로 데이터 기반의 예측 모델을 개발하는 것을 목표로 합니다. 예측된 복구 금액을 통해 향후 수해 피해 예방을 위한 비용 예산 편성에 참고할 수 있도록 하는 것이 주요 목적입니다.

- 분석도구

SQL, Python, Power BI를 활용하여 장마기간 강수량과 수해 피해 간의 관계를 분석하고, 데이터 기반 예측 모델을 통해 피해 최소화 방안을 모색합니다.

2. 데이터 수집 및 전처리

- 데이터 출처(수집)

  - 국민재난안전포털(호우 피해정보, 위험요인 정보)

  https://www.safekorea.go.kr/idsiSFK/neo/main/main.html

- 기상청 기상자료개방포털(강수량 정보, 지점 정보)
  https://data.kma.go.kr/cmmn/main.do;jsessionid=fna6MAZkfb1JWAEvwahpeyOTsgUwEoWhEL3001gJ0HAEWFx0z8ytF5813pWVqTqp.was02_servlet_engine5

- 행정안전부 (재해연보)
  https://www.mois.go.kr/frt/bbs/type001/commonSelectBoardArticle.do;jsessionid=HiilyP2FySqoKhg-mRt6-pFA.node30?bbsId=BBSMSTR_000000000014&nttId=97685

2. 데이터 수집 및 전처리

• 데이터 설명

연도 : 2012년 ~2021년 10년간 데이터

시도 : 전국 16개 시도(세종시는 측정지점이 없어 제외)

수집할 데이터는 아래와 같이 정리하였습니다.

테이블명	PK	PK								
장마 강수량 테이블	시도코드	연도	시도명	지점코드	지점명	시작일	종료일	장마일수	합계강수량	
피해금액 및 복구비용 테이블	시도코드	연도	시도명	건물	선박	농경지	공공시설	사유시설	복구비	
위험도 테이블	시도코드		시도명	인구	가구	주택	노후건물	지하건물	시설	농경지면적
지점정보	시도코드		시도명	지점코드	지점명	측정시작일	측정종료일			
행정구역코드	시도코드		시도명							

2. 데이터 수집 및 전처리

- 데이터 전처리

1. 수집된 데이터를 csv형식으로 변경
2. 집계 기준을 연도/시도별로 통일
3. 테이블 결합을 위한 FK코드 부여
4. 위험요인 데이터 스케일링(인구, 세대, 면적 등 수치 기준이 모두 다름)

Ex) 장마 강수량 정보는 관측지점을 기준으로 측정하고 피해정보는 시도를 기준으로 산정하므로 관측지점의 주소정보를 결합하여 시도별로 집계한 후 피해금액과 결합

테이블 ERD

1. 시도코드

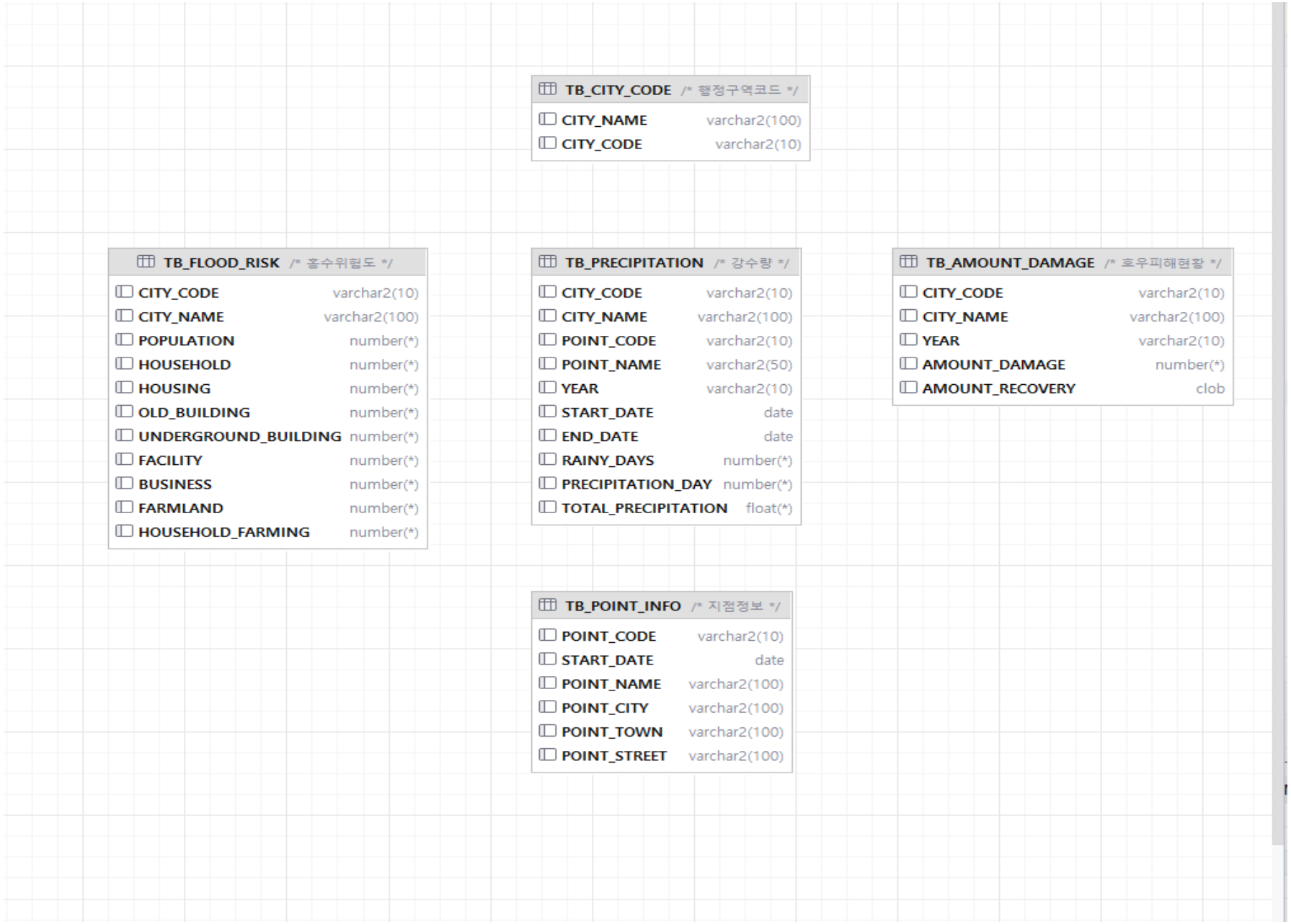
2. 위험요인

- 시도코드, 시도명, 인구, 가구, 주택, 노후건물, 지하건물, 공공시설, 사업체, 농지, 농림어업가구

3. 장마 강수량

4. 피해 및 복구금액

5. 지점정보



3. 탐색적 데이터분석

- 연도별 강수량 및 피해금액 데이터

연도	총 강수량(mm)	강수일수(일)	장마기간(일)	피해금액(원)	복구금액(원)
2012	1,254	16	25	2,398,690,188	6,178,630,625
2013	1,685	26	47	9,883,050,125	23,604,596,250
2014	668	16	29	8,888,215,875	29,098,092,500
2015	1,046	19	35	75,791,500	101,315,125
2016	1,408	18	33	2,242,925,375	2,300,559,313
2017	1,189	18	30	6,347,457,313	19,921,660,563
2018	1,196	12	15	3,315,077,750	7,593,758,438
2019	1,273	17	33	103,008,063	335,634,750
2020	2,847	30	46	68,372,617,875	221,264,060,750
2021	919	11	17	2,540,375,250	11,759,592,750
평균	1,348	18	31	10,416,720,931	32,215,790,106

3. 탐색적 데이터분석

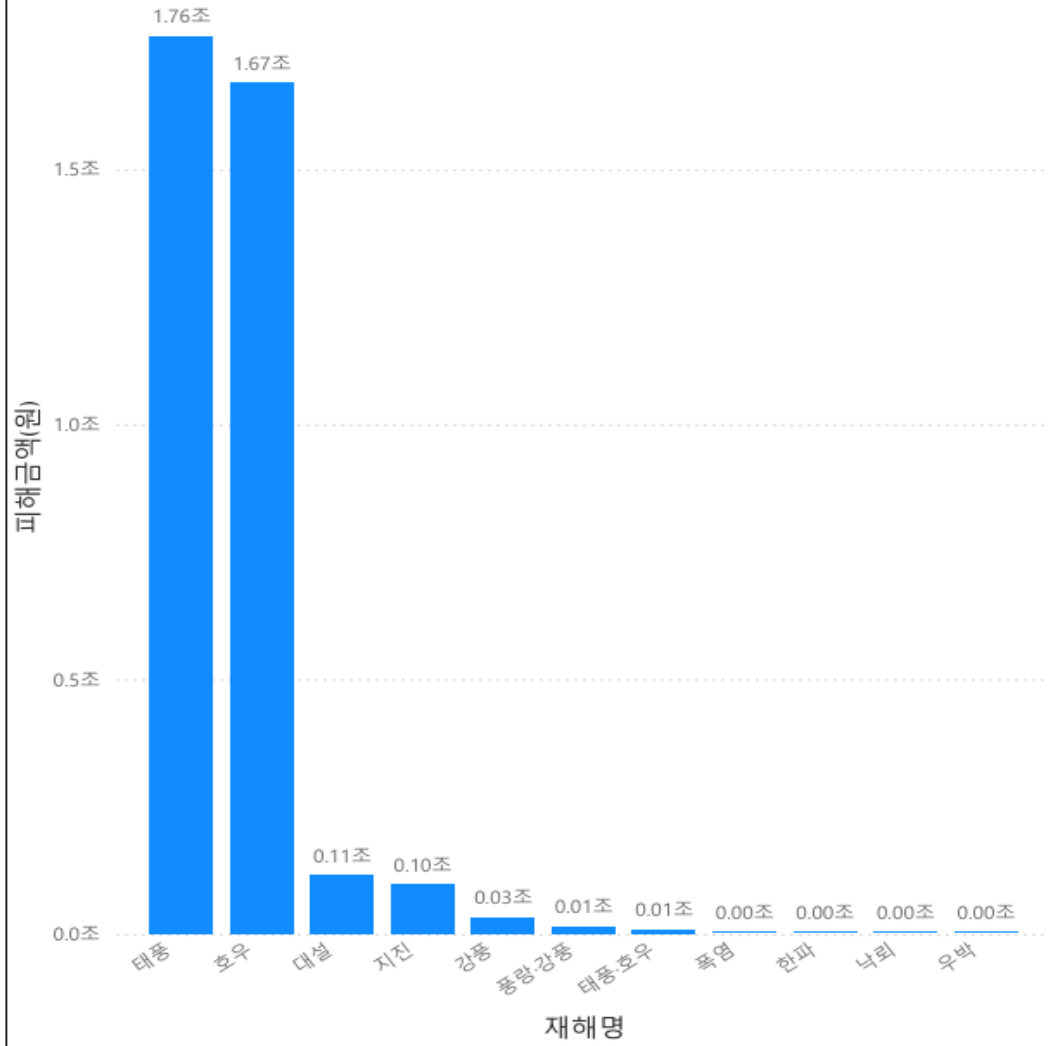
- 위험요인 데이터

시도명	인구	세대	주택	노후건물	지하건물	공공시설	사업체	농지(km^2)	농림어업 세대
서울특별시	8,981,223	4,096,251	3,295,453	319,272	372,249	9,914	1,093,498	7	9,652
부산광역시	3,211,947	1,445,879	1,295,883	207,206	59,909	2,711	380,409	53	15,203
대구광역시	2,322,535	1,019,724	825,931	147,933	43,269	2,445	263,233	141	32,038
인천광역시	2,849,377	1,210,429	1,123,913	96,319	69,880	3,574	293,139	187	15,306
광주광역시	1,422,849	623,128	536,025	110,653	22,263	1,791	161,268	92	18,448
대전광역시	1,429,165	644,559	495,870	75,853	43,688	2,307	154,532	41	15,310
울산광역시	1,075,803	454,357	385,691	61,745	14,041	1,282	110,898	92	16,819
세종특별자치시	371,526	153,546	141,122	11,156	4,178	543	32,119	76	7,814
경기도	13,013,766	5,412,106	4,817,139	324,304	280,690	20,053	1,441,118	1,542	128,191
강원특별자치도	1,483,548	684,720	603,698	170,128	43,241	2,152	196,224	1,029	77,848
충청북도	1,559,996	708,443	614,805	187,738	35,129	2,121	189,216	979	73,968
충청남도	2,067,057	930,938	832,488	215,582	32,493	3,325	256,612	2,165	138,193
전북특별자치도	1,714,814	777,801	686,137	209,015	24,237	2,482	223,703	1,920	106,664
전라남도	1,707,852	784,792	734,210	361,537	18,594	2,153	232,526	2,678	167,417
경상북도	2,507,427	1,157,514	1,009,342	393,377	29,232	3,577	316,532	2,278	184,287
경상남도	3,159,685	1,394,074	1,228,453	352,227	44,425	5,735	382,347	1,308	143,084
제주특별자치도	643,966	276,225	240,849	87,961	17,499	886	91,636	550	34,432

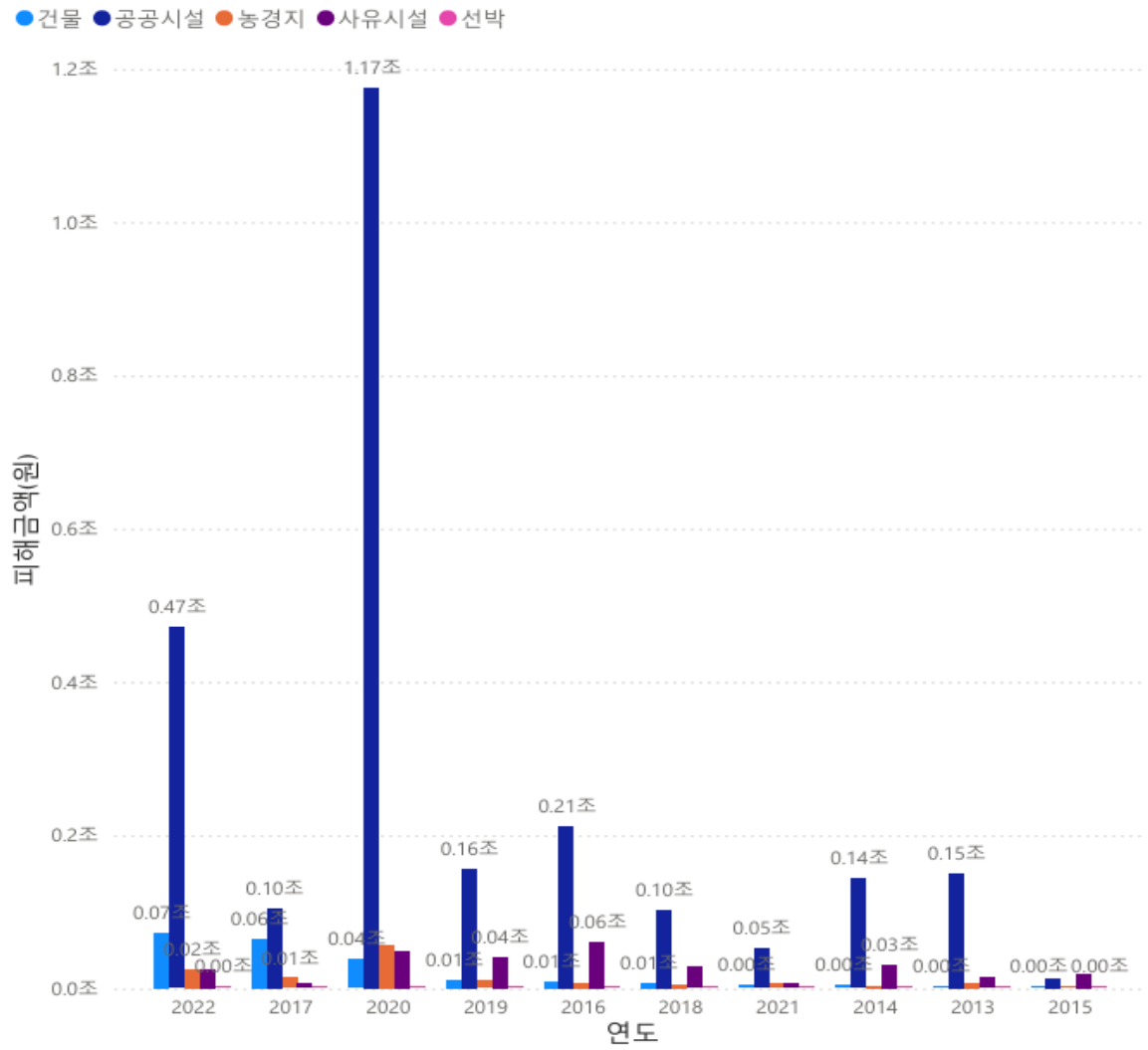
3. 탐색적 데이터분석(시각화)

- 호우피해는 재해유형 중 2위, 시설피해는 공공시설이 1위

재해유형별 피해현황

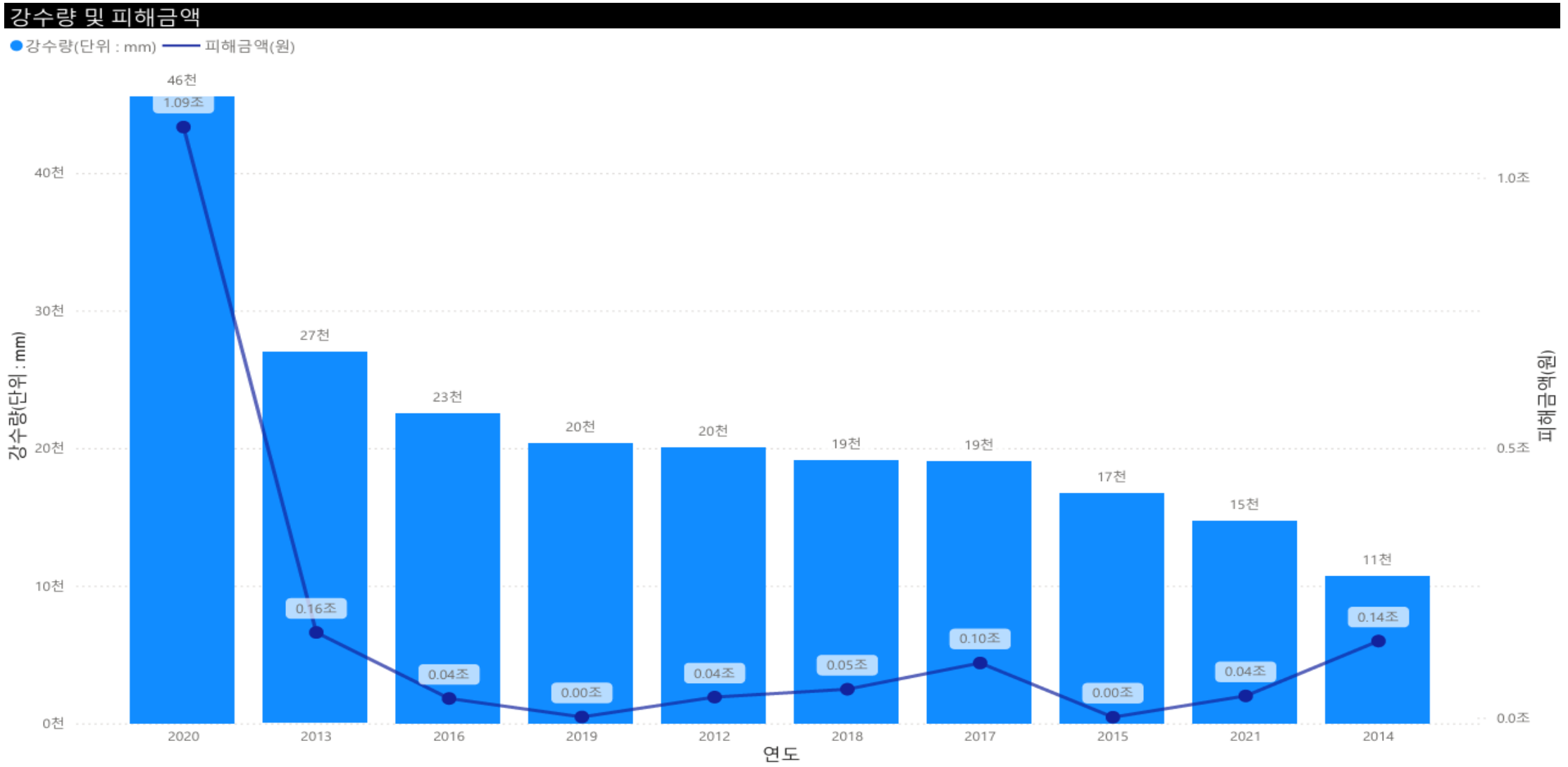


시설별 피해현황



3. 탐색적 데이터분석(시각화)

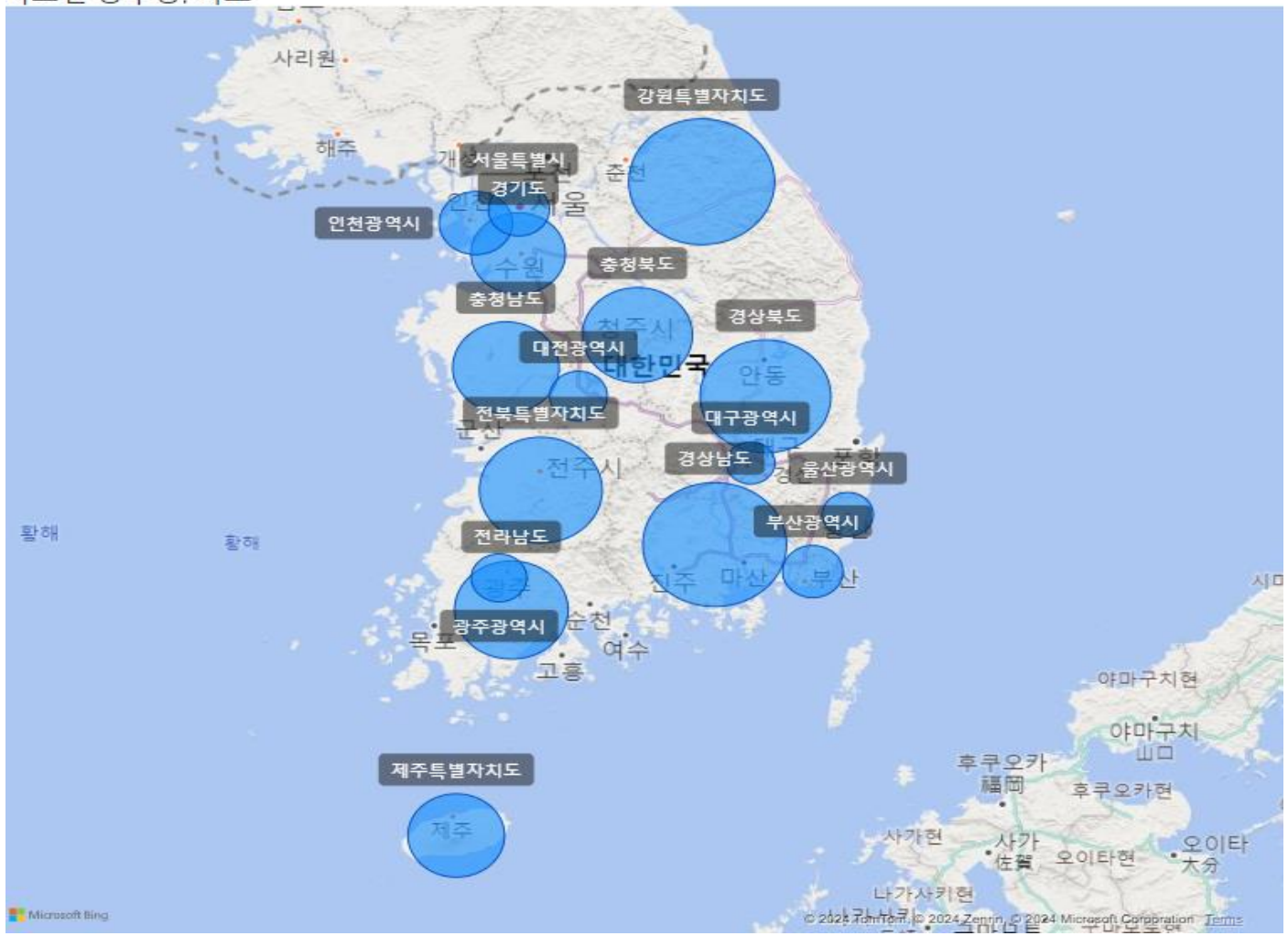
- 강수량 및 피해금액은 2020년이 가장 높음.



3. 탐색적 데이터분석(시각화)

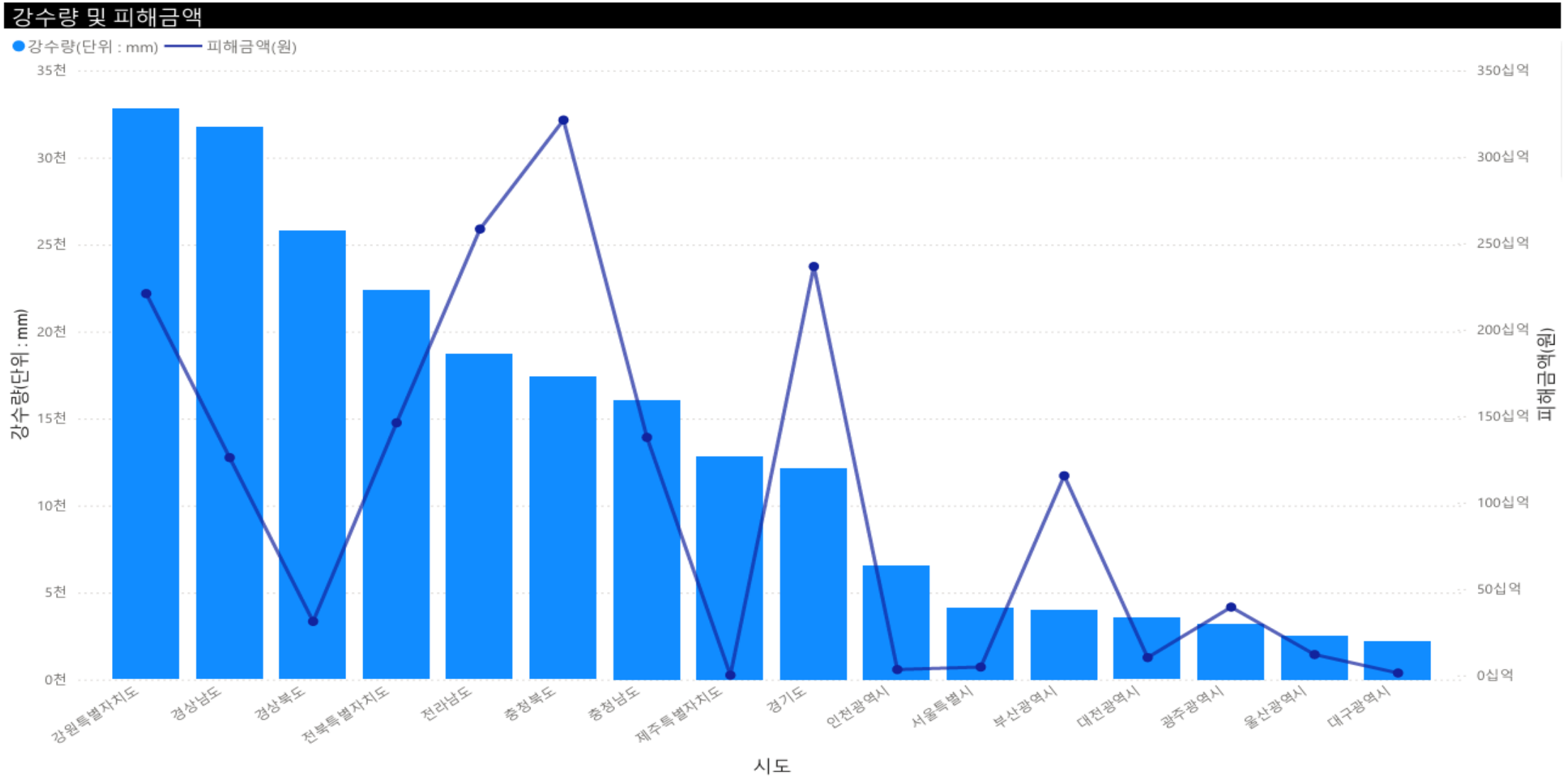
- 강수량은 해안 지역이 높고 내륙 지역은 낮다.

시도별 강수량, 시도



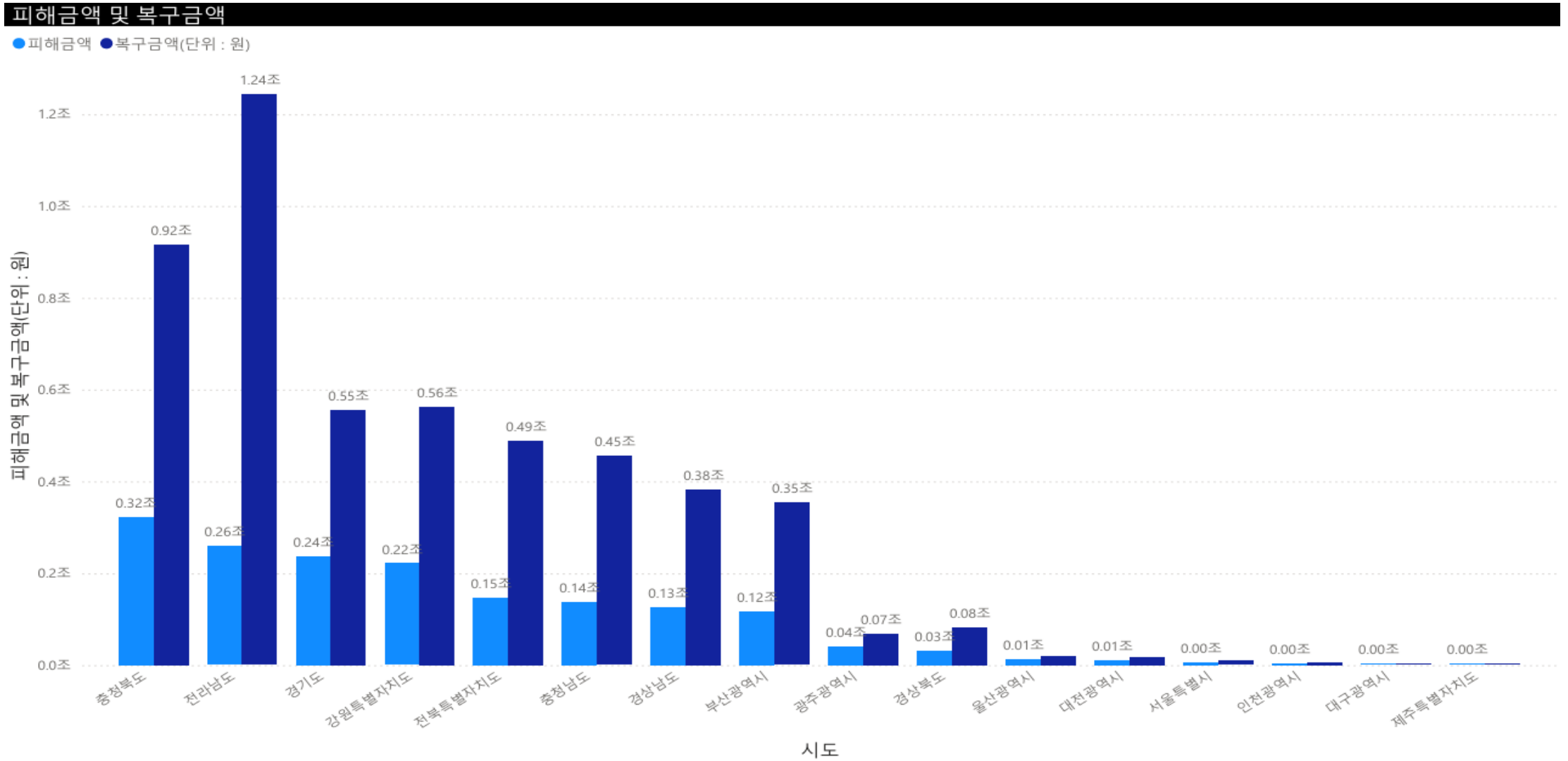
3. 탐색적 데이터분석(시각화)

- 충청북도, 전라남도, 경기도는 강수량 대비 피해금액이 높다.



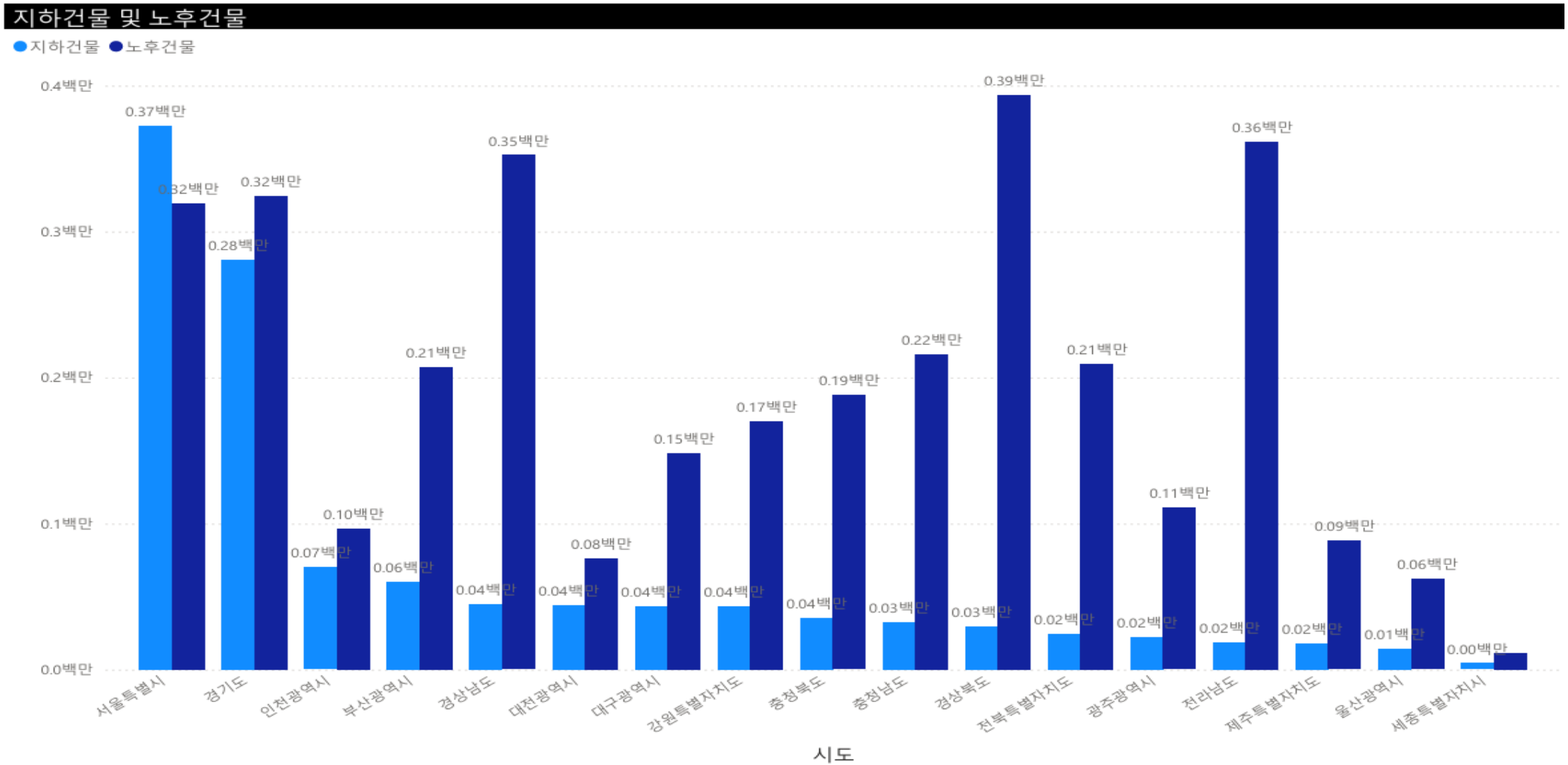
3. 탐색적 데이터분석(시각화)

- 피해금액 대비 복구비용은 충청북도, 전라남도가 높다.



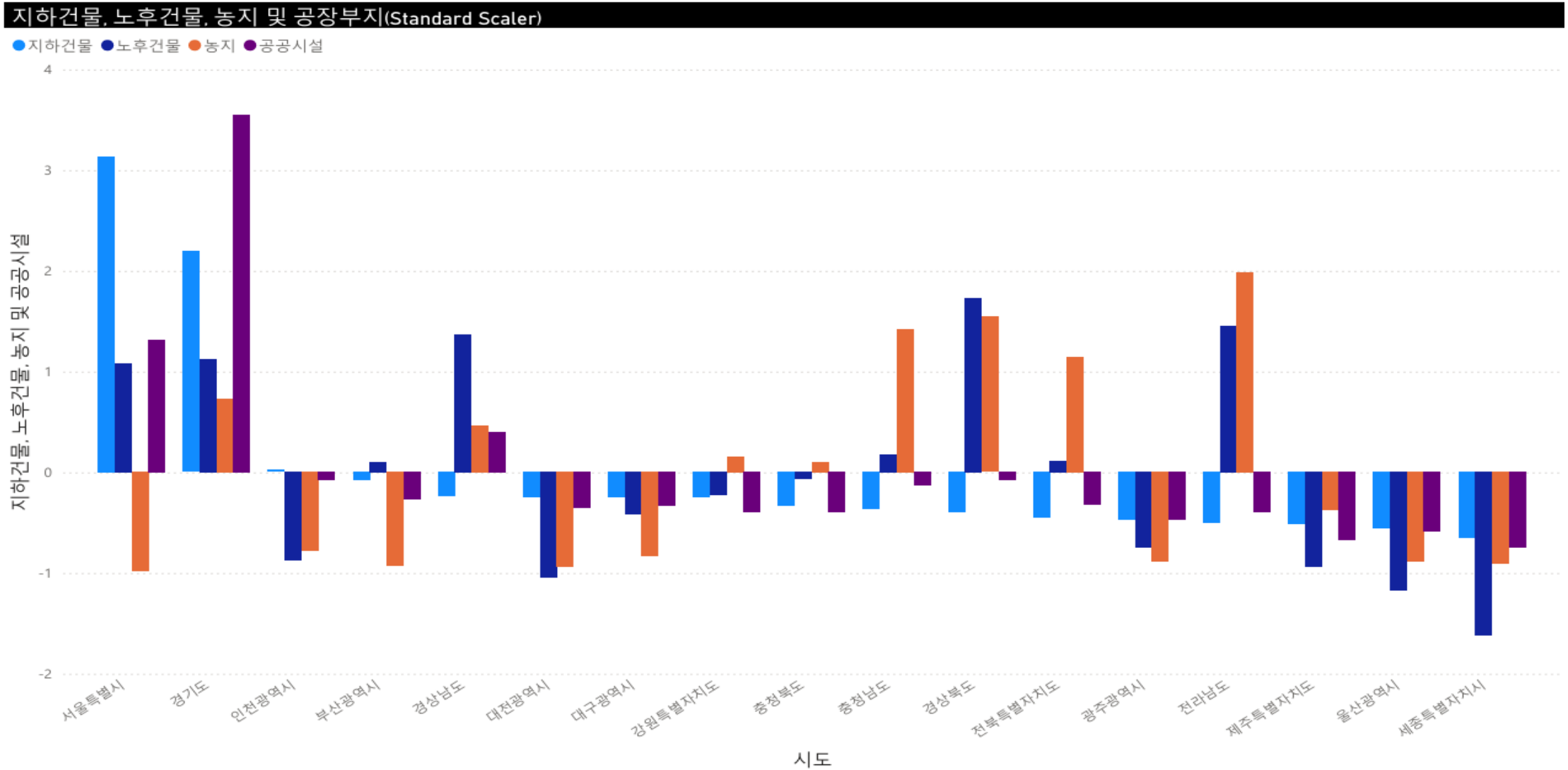
3. 탐색적 데이터분석(시각화)

- 위험요인 중 지하건물은 수도권에 집중 되어있고 노후건물은 경상도, 전라남도가 많음.



3. 탐색적 데이터분석(시각화)

- 지하건물과 공공시설은 수도권에 집중, 지방은 노후건물과 농지 비율이 높다.



3. 탐색적 데이터분석

- 탐색요약

1. 장미기간은 6월 말 ~ 7월 말로 기간은 평균 31일, 강수일수는 평균 18일

2. 연간 평균 강수량은 1,348mm 이고 2020년은 2,847mm로 강수량이 가장 높았다.

3. 강수량이 가장 높은 지역은 강원도, 경상도, 전라도 순으로 해안지역이다.

4. 충청북도, 전라남도, 경기도는 강수량 대비 피해금액이 높은 취약지역이다.

5. 경상북도는 강수량에 비해 피해규모가 낮다.

4. 회귀분석

- 회귀분석

1. 2012년~2020년까지의 데이터를 학습하여 2021년 복구금액 예측
2. 독립변수 : 총 강수량, 강수일수, 장마기간, 피해금액, 복구금액, 인구, 세대, 주택, 노후건물, 지하건물, 공공시설, 사업체, 농지, 농림어업 세대

   종속변수 : 복구금액
3. 데이터 전처리 : 데이터 스케일, 각 연도별 데이터에 위험요인 데이터 Merge,
   불필요한 열 제거(시도코드, 측정 값이 없는 세종시, Object 데이터 등)
4. 모델 : 해당 데이터는 시계열 데이터로 회귀 모델 Linear, Ridge, Lasso 적용.

4. 회귀분석

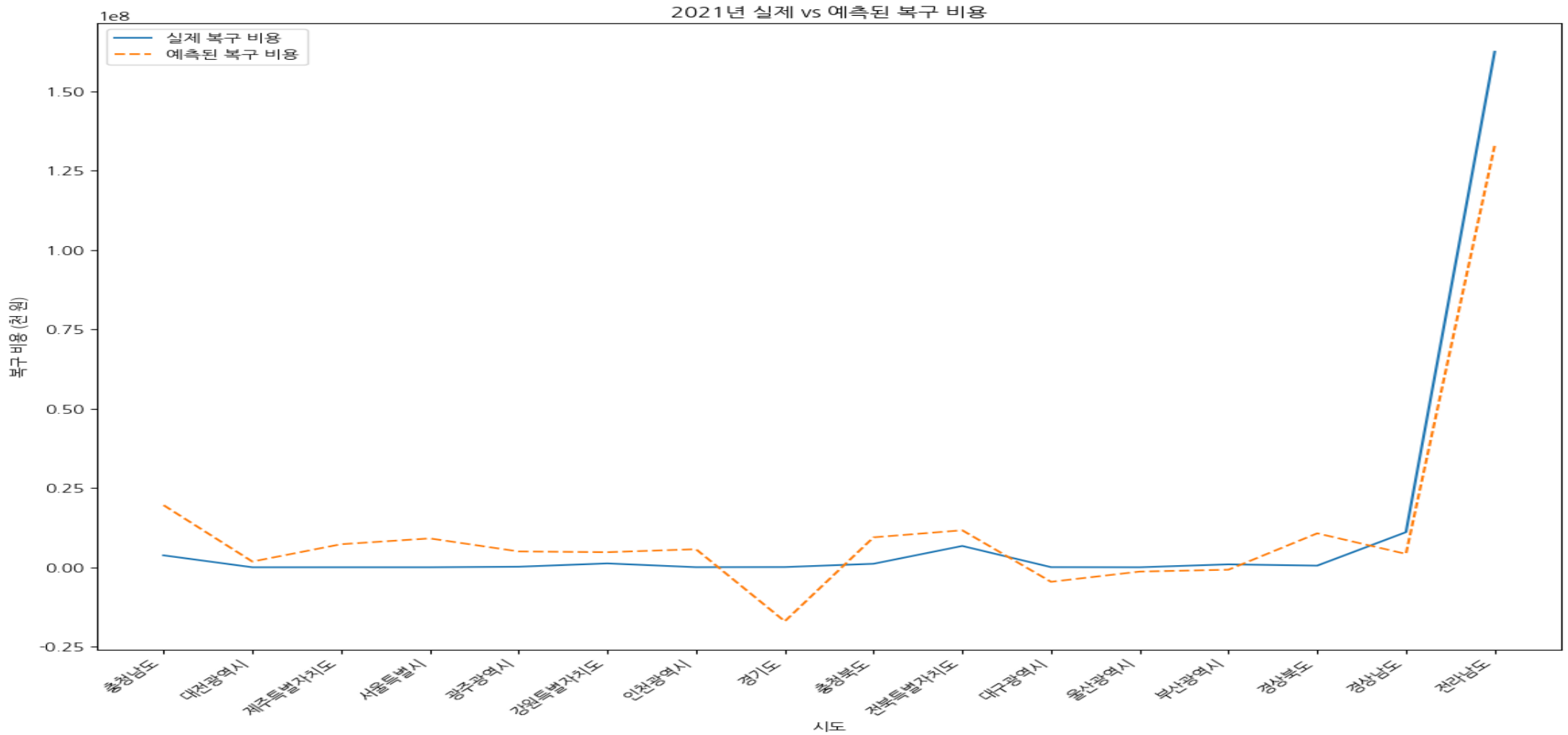
- 학습결과

모델명	MSE	MAE	R-squared
Linear	0.01	0.07	0.92
Ridge	0.01	0.07	0.89
Lasso	0.02	0.04	0.80

- Mean Squared Error (MSE): 예측값과 실제값의 오차 제곱의 평균입니다.
- Mean Absolute Error (MAE): 예측값과 실제값의 오차의 절대값 평균으로, MSE보다 해석이 쉽고, 이상치의 영향을 덜 받습니다.
- R-squared (R^2): 모델이 데이터를 얼마나 잘 설명하는지를 나타내는 지표로, 1에 가까울수록 모델이 잘 설명하고 있음을 의미합니다.

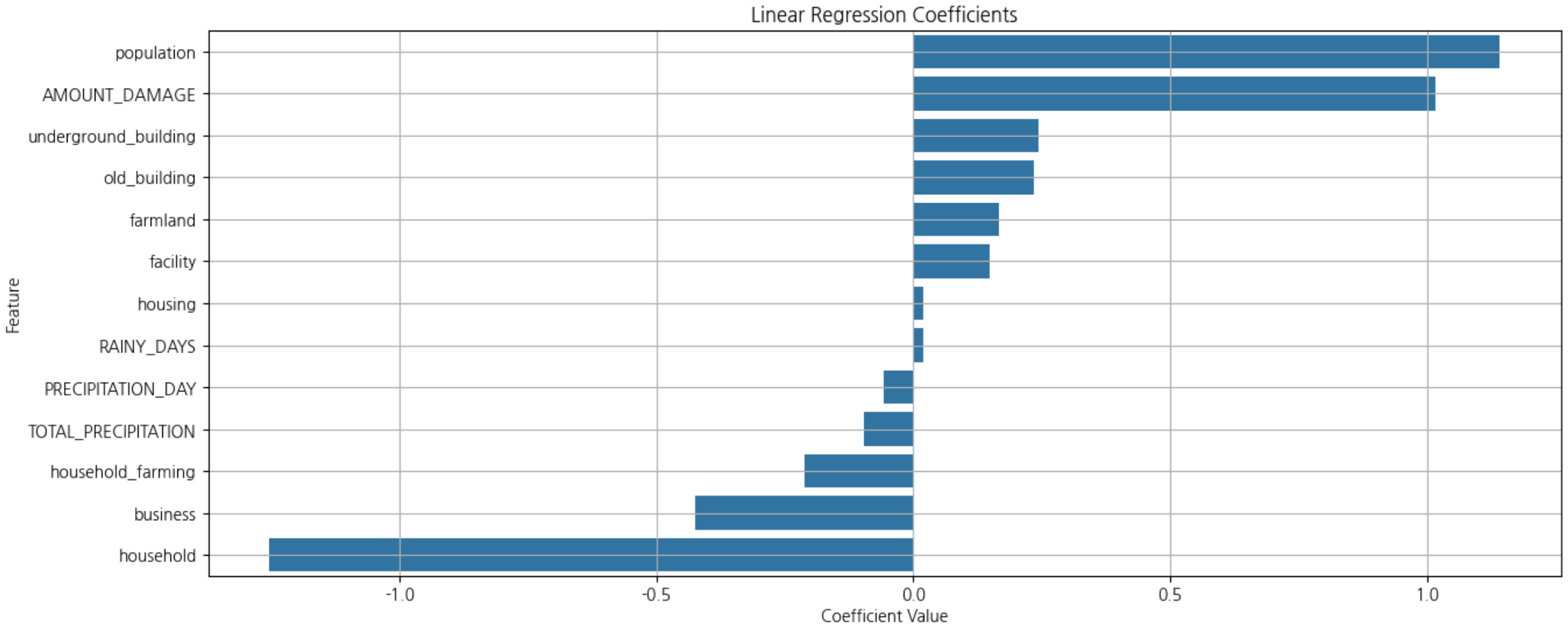
4. 회귀분석

- 예측결과(Linear 모델)



4. 회귀분석

- 변수별 영향 계수



4. 회귀분석

- 변수별 영향 계수

특성	계수(양)
인구	1.141756
피해금액	1.016551
지하건물	0.243807
노후건물	0.234763
농지	0.168065
공공시설	0.149304
주택	0.019772
강수일자	0.01895

특성	계수(음)
장마기간	-0.056964
총 강수량	-0.095314
농림어업세대	-0.211726
사업체	-0.424111
세대	-1.252593

5. 결과 및 해석

- 분석 결과

1. 분석 모델은 복구비용을 92%의 정확도로 예측.

2. 충청북도와 전라남도는 호우피해 취약 지역.

3. 경상북도는 강수량에 비해 피해가 낮은 지역.

4. 호우 피해는 피해금액 대비 복구금액이 3배 높다.

5. 강수량보다는 지역의 특성에 따라 복구금액에 차이가 나타난다.

5. 결과 및 해석

- 한계

1. 정확도 한계 : 향후 피해를 예측하기 위해 강수량과 피해금액을 독립변수로 사용해야 하는데, 이 값들이 예측치에 의존하기 때문에 정확도에 따라 모델의 신뢰도가 낮아질 수 있습니다.
2. 재난예방 예산편성의 제한 : 이번 분석은 시도를 기준으로 취약지역을 파악하여 해당 지역에 맞는 예산 편성을 지원하기 위한 것입니다. 그러나 실제 재난예방 예산은 사업을 확정하고 업체를 선정하는 방식으로 편성되므로, 시도별 지원과는 차이가 있습니다. 다만, 업체에서 분석을 참고하여 취약지역 중심으로 예방 사업을 수행하는 데 도움이 될 수 있습니다.
3. 인명피해와 이재민 피해의 제한 : 분석에는 인명피해(사망, 실종)와 이재민 피해처럼 금액으로 환산하기 어려운 피해 요소가 포함되지 않았습니다. 따라서 실제 피해 규모는 분석 결과보다 클 수 있습니다.

6. 결론

- 결론

이번 분석은 호우 피해 예방에 기여하기에는 한계가 있지만, 복구비용을 예측함으로써 보다 적절한 예방 예산을 편성하여 피해를 줄이는 데 도움이 되기를 바랍니다.

특히, 인명피해와 이재민들의 고통은 수치로 환산할 수 없는 만큼, 예방의 중요성은 더욱 강조되어야 합니다.

앞으로도 호우 피해에 대한 지속적인 관심과 예방 노력이 이어져 더 많은 사람들이 안전하게 일상을 지킬 수 있는 환경이 조성되기를 기대합니다.

7. 부록

- 코드 경로

- https://github.com/jungseungtae/portpolio/blob/master/Portfolio/python/rainfall.py

끝.