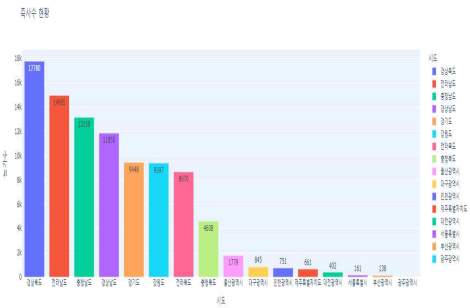
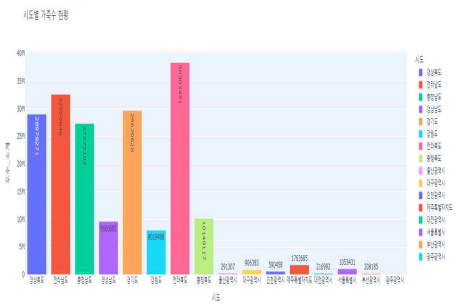
붙임 2-3 **대국민 분석과제 아이디어 기획서(서식)**

2022년도 환경데이터 분석 · 활용 공모전(아이디어 분야)	
팀 명	최강이지~
제안자	① 이준형 ② 배지혜 ③ 최솔 ④ 강지아
아이디어명	회귀모델을 활용한 가축 환경오염 지수 시각화
과제 번호	2 쾌적하고 안전한 환경
분야	☑기후대기 ☑물환경 ☐ 수자원 ☐ 상하수도 ☐ 자연환경 ☐ 자원순환 ☐ 환경보건 ☐ 화학물질 ☐ 환경산업 ☑생활환경
요약	단계적 선택법을 활용한 최적의 회귀모델을 구축하고, 회귀계수를 기반으로 시도, 시군별 가축 분뇨로 발생하는 환경오염도 지수를 산출해 QGIS에서 지도 시각화를 진행하였다.
필요성 및 목적	가축 분뇨는 토양을 비옥하게 하고, 친환경적인 먹거리를 생산하는 순기능이 존재한다. 그러나, 현재 가축사육 규모가 커지면서 지역의 자연환경이 수용하기 어려울 정도로 많은 양의 분뇨가 발생하고 있다. 통제되지 않는 상황에서 농경지에 퇴비, 액비의 무분별한 살포는 필요 이상의 양분(질소, 인 등)이 토양에 과잉으로 축적되어 강우 시 지하수 및 주변 하천으로 유입되어 수질을 악화시킬 우려가 매우 크다. 특히, 악취를 동반하는 암모니아는 토양 및 수질 오염은 물론, 자동차 등에서 나오는 가스상 물질과 결합해 2차 미세먼지를 만들어 낼 수 있어 전반적으로 환경문제에 치명적인 영향을 줄 것으로 판단하였다. ☐ 2020년도 시도별 축사 현황



[그림 1-2] 시도별 축사 현황



[그림 1-1] 시도별 가축수 현황

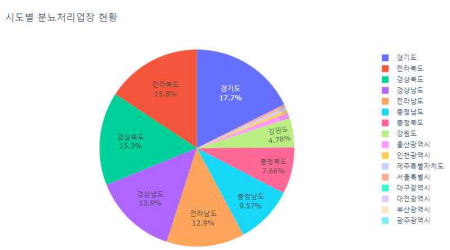
시도별 축사 현황을 시각화한 결과, 광역시보다 지방에서의 축사 수가 많음을 파악할 수 있다. 축사 수가 많은 시도가 대체로 높은 가축 수를 보이고 있으며, 경기도와 전라북도는 축사 수 대비 가축 수가 많으므로 공장식 농장의 형태로 운영한다는 것을 추측할 수 있다. 공장식 농장은 축사 면적 대비 가축의 밀도가 높은 것을 말하며, 좁은 공간에서 더 많은 분뇨를 발생시켜 해당 지역의 환경오염을 초래할 수 있으므로 더욱 주의해서 살펴볼 필요가 있다.

□ 2020년도 시도별 환경오염 현황

[그림 1-3] 시도별 수질오염도

[그림 1-4] 시도별 토양오염도

[그림 1-5] 시도별 대기오염도

	하루 평균 쓰레기 배출량이 가장 높은 서울특별시를 제외하고, 축사 수가 많은 지역에서 수질, 대기, 오염도의 수치가 높음을 파악할 수 있다. 가축 분뇨발생량에 따른 하수처리 시설 및 분뇨처리량 시설의 현황을 파악해 지역별 보충이 필요한 시설을 확인해 보고자 한다. □ 2020년도 시도별 처리시설 현황  [그림 1-7] 시도별 분뇨처리업장 현황  [그림 1-6] 시도별 하수처리시설 현황 분뇨 처리 업장 같은 경우, 축사 수가 세 번째로 많은 충청남도가 다른 시도에 비해 상대적으로 적은 분뇨처리 시설 개수가 적다. 또한, 축사 수가 상위권에 존재하는 강원도도 분뇨처리 시설의 비율이 낮음을 파악할 수 있다. 하수처리시설의 경우, 축사와 가축 수가 많고, 수질 오염 현황에서 가장 높은 수치를 보인 경기도는 다른 지역에 비해 하수처리시설이 적은 비율로 차지하고 있다. 전반적으로 축사 수 및 가축 수가 많거나, 축사 수 대비 가축 수의 밀집도가 높은 지역에 대한 수질, 토양, 대기 오염도가 높다는 것을 시각화를 통해 나타냈으며, 축산업으로 발생 되는 가축 분뇨가 환경오염에 영향을 미친다는 것을 파악할 수 있었다. 지역이 소화할 수 있는 양분의 처리량과 지역별 축사 규모 기준에 관련한 정책을 마련해야 할 필요가 있고, 이에 본 팀은 전국 시군구를 대상으로 회귀분석 모델을 만들어 도출된 회귀계수를 가중치로 활용해 해당 연도의 환경오염지수를 개발할 것이다.
과제정의	시군구별 가축 환경과 인구수, 날씨 등 지역별 특성을 반영한 회귀모델을 구성하고, 가축분뇨로 인한 환경오염지수를 개발하여 위험구역을 발굴해 쾌적하고 안전한 환경을 위한 정책 수립에 보탬이 되고자 한다.
활용데이터	***모든 데이터는 무료로 제공받음** **환경부** -2020년도 기준 광역지자체별 가축분뇨 처리량.xlsx -2020년도 기준 광역지자체별 가축분뇨 처리농가수.xlsx

	-2020년도 기준 광역지자체별 가축분뇨 발생량.xlsx -2020년도 기준 광역지자체별 가축사육 농가수 및 두수.xlsx **물 환경 정보 시스템** -2020년 질산성질소, 암모니아성질소, 총대장균수, 군원성대장균 총 4종의 오염물질 크롤링 진행 -2020년도 시도별 하천 검색을 통해 결측치 채움 **공공데이터포털** -fulldata_02_04_01_P_가축사육업.csv(전국 가축업 현황) -fulldata_09_30_01_P_가축분뇨수집운반업.csv(전국 가축분뇨수집운반업 현황) -한국환경공단_공공하수처리시설 현황_20201231.csv(전국 하수처리 시설 현황) **기상청** -기후통계분석-> 기상현상일수 -> 폭염일수에서 2020년도 전국 폭염일수 csv파일 다운로드 **농림축산식품부, 국가가축방역통합시스템** -농림축산식품부의 가축질병 발생현황 페이지에서 아프리카돼지열병, 조류인플루엔자, 구제역 데이터 크롤링 진행 -국가가축방역통합시스템에서 질병 발생 농가 데이터 크롤링 진행 **오픈마켓** -전국 읍면동 shp파일 다운로드 -> QGIS에 업로드하여 전국 읍면동별 면적 도출 -2020년도 전국 읍면동 면적 데이터 엑셀 파일 다운로드 **행정안전부** -202012_202012_주민등록인구 및 세대현황_연간.xlsx(2020년도 전국 읍면동별 거주자 인구 등록 현황 데이터 다운로드) KOSIS

	-시도, 시군별 1~12월 대기오염 지수 데이터 엑셀 파일 다운로드 -2020년도 읍면동 인구 데이터 엑셀 파일 다운로드 -2020년도 가축 종사자 수 데이터 엑셀 파일 다운로드 -2020년도 시도, 시군별 강수량 데이터 엑셀 파일 다운로드 -시군, 읍면동별 면적 데이터 엑셀 파일 다운로드 **토양지하수정보시스템, 한국환경정책평가연구원** -2020년도 시도, 시군, 읍면동을 기준으로 전국 토양 오염도 데이터 다운로드 AirKorea -2020년도 대기오염도 결측치 채울 때 참고한 사이트
분석방법	1. 사이킷런의 IterativeImpute 모델을 활용하여 결측치를 처리한다. 2. 변수의 정규성 및 단위 통합을 위한 로그 변환, Min-Max 스케일링을 수행한다. 3. 변수 선택법 방법 중, 단계적 선택법을 활용해 최적의 회귀모델을 선별한다. ※ 선별된 변수: 시도별_가축더위지수, 읍면동_총거주자수, 가축분뇨발생량_합계_면적비, 가축사육종사자수, 분뇨처리업장_개수, 분뇨처리_정화 4. 선별된 모델에 존재하는 변수의 회귀계수를 각 변수의 가중치로 적용해 시군구별 가축분뇨로 발생 되는 환경오염도 지수를 산출한다. **지수 산출식 = ∑해당 변수의 회귀계수(가중치) X 해당 변수값** 5. 산출된 지수를 기반으로 QGIS에서 지도 시각화를 수행하여 오염도 지수가 높은 위험구역을 발굴한다.
기대효과	축산업 현황, 지역별 특성 그리고 토양, 대기, 수질 데이터를 기반으로 시군구별 축사로 인한 환경오염도 지수로 두 가지 측면에서의 효과를 기대할 수 있다. 첫째, 정책을 제정하는 기관에서 환경오염도 지수를 바탕으로 가축분뇨 공공 처리시설이 필요한 지역의 우선순위 선정과 설치 근거로 활용될 수 있다.

	둘째, 축사 수 대비 가축 수 및 분뇨발생량의 밀도가 높은 지역에 대한 축사 면적, 가축 수 제한 등 규제를 확립하기 위한 근거에 보탬이 될 수 있다.

※ 분량은 최대 10페이지 이내, 글자체 및 글자크기는 서식과 동일하게 작성

※ 과제와 관계없는 내용 또는 참가자의 인적 사항을 기재하면 자동 실격 처리

첨부1 데이터 전처리

분석방법1. 결측치 처리

```
from sklearn.experimental import enable_iterative_imputer
from sklearn.impute import IterativeImputer

imputer = IterativeImputer(max_iter = 10, random_state = 0)
df_num = pd.DataFrame(imputer.fit_transform(df_num))
df_num.columns = df_num_col
df_num
```

IterativeImputer 모델을 활용해 변수의 결측치 처리 수행하였다.

분석방법2. 정규화 및 표준화 수행

```
## numerical 변수를 정규분포에 가깝게 통계적 변환
df['폭염일수_2020'] = np.log1p(df['폭염일수_2020'])
df['가축분뇨발생량_합계'] = np.log1p(df['가축분뇨발생량_합계'])
df['분뇨처리_총합'] = np.log1p(df['분뇨처리_총합'])
df['분뇨처리_정화'] = np.log1p(df['분뇨처리_정화'])
df['분뇨처리_퇴비'] = np.log1p(df['분뇨처리_퇴비'])
df['읍면동_면적'] = np.log1p(df['읍면동_면적'])
df['하수처리시설_개수'] = np.log1p(df['하수처리시설_개수'])
df['분뇨처리업장_개수'] = np.log1p(df['분뇨처리업장_개수'])
df['읍면동_총거주자수'] = np.log1p(df['읍면동_총거주자수'])
df['질병발생'] = np.log1p(df['질병발생'])
df['시도별_가축더위지수'] = np.log1p(df['시도별_가축더위지수'])
df['강수량_2020'] = np.log1p(df['강수량_2020'])
df['가축사육종사자수'] = np.log1p(df['가축사육종사자수'])
df['두수_한우'] = np.log1p(df['두수_한우'])
df['두수_돼지'] = np.log1p(df['두수_돼지'])
df['두수_말'] = np.log1p(df['두수_말'])
df['두수_닭_오리'] = np.log1p(df['두수_닭_오리'])
df['두수_소계'] = np.log1p(df['두수_소계'])
df['농가수'] = np.log1p(df['농가수'])
df['농가수_면적비'] = np.log1p(df['농가수_면적비'])
df['가축분뇨발생량_합계_면적비'] = np.log1p(df['가축분뇨발생량_합계_면적비'])
df['읍면동_총거주자수_면적비'] = np.log1p(df['읍면동_총거주자수_면적비'])
df['두수_소계_면적비'] = np.log1p(df['두수_소계_면적비'])
df['시군별_수질오염도'] = np.log1p(df['시군별_수질오염도'])
df['시군별_대기오염도'] = np.log1p(df['시군별_대기오염도'])
df['시군별_토양오염도'] = np.log1p(df['시군별_토양오염도'])

## 표준화
from sklearn.preprocessing import MinMaxScaler

min_max = MinMaxScaler()
df_num1 = pd.DataFrame(min_max.fit_transform(df_num))

col = df_num.columns.tolist()
df_num1.columns = col
df_num1
```

로그변환을 통해 변수변환을 수행, Min-Max 스케일링을 수행해 변수들의 정규성및 단위 통일을 진행하였다.

첨부2 변수선택법을 활용한 회귀모델 선별

분석방법3. 단계적 선택법으로 최적의 회귀모델 선정

OLS Regression Results			
Dep. Variable:	sum	R-squared:	0.533
Model:	OLS	Adj. R-squared:	0.519
Method:	Least Squares	F-statistic:	37.66
No. Observations:	205	Prob (F-statistic):	2.69e-30
Df Residuals:	198	Log-Likelihood:	-33.389
Df Model:	6	AIC:	80.78
Covariance Type:	nonrobust	BIC:	104.0

	coef	std err	t	P>|t|	[0.025	0.975]
const	31.9591	4.957	6.447	0.000	22.184	41.734
시도별_가축더위지수	-7.6110	1.172	-6.494	0.000	-9.922	-5.300
읍면동_총거주자수	0.1391	0.025	5.528	0.000	0.090	0.189
가축분뇨발생량_합계_면적비	8.569e+04	2.08e+04	4.129	0.000	4.48e+04	1.27e+05
가축사육종사자수	-0.0834	0.019	-4.381	0.000	-0.121	-0.046
분뇨처리업장_개수	0.1123	0.041	2.751	0.006	0.032	0.193
분뇨처리_정화	0.0240	0.012	2.031	0.044	0.001	0.047

단계적 선택법 수행 결과, 총 6개의 변수로 구성된 모델이 선택되었으며 모든 변수는 P-value 값이 제1종 오류인 0.05보다 작아 귀무가설을 기각하여 모두 유의미하다는 결과를 보였다. 또한, 사회과학에서 통상적으로 유의미하다고 판단하는 R^2은 0.4로, 해당 수치를 넘었기 때문에 설명력 있는 모델이라 할 수 있다.

선택된 변수의 회귀계수를 가중치로 이용하여 시군구별 환경오염도 지수를 산출하고자 한다.

첨부3 QGIS를 활용한 환경오염지수 시각화

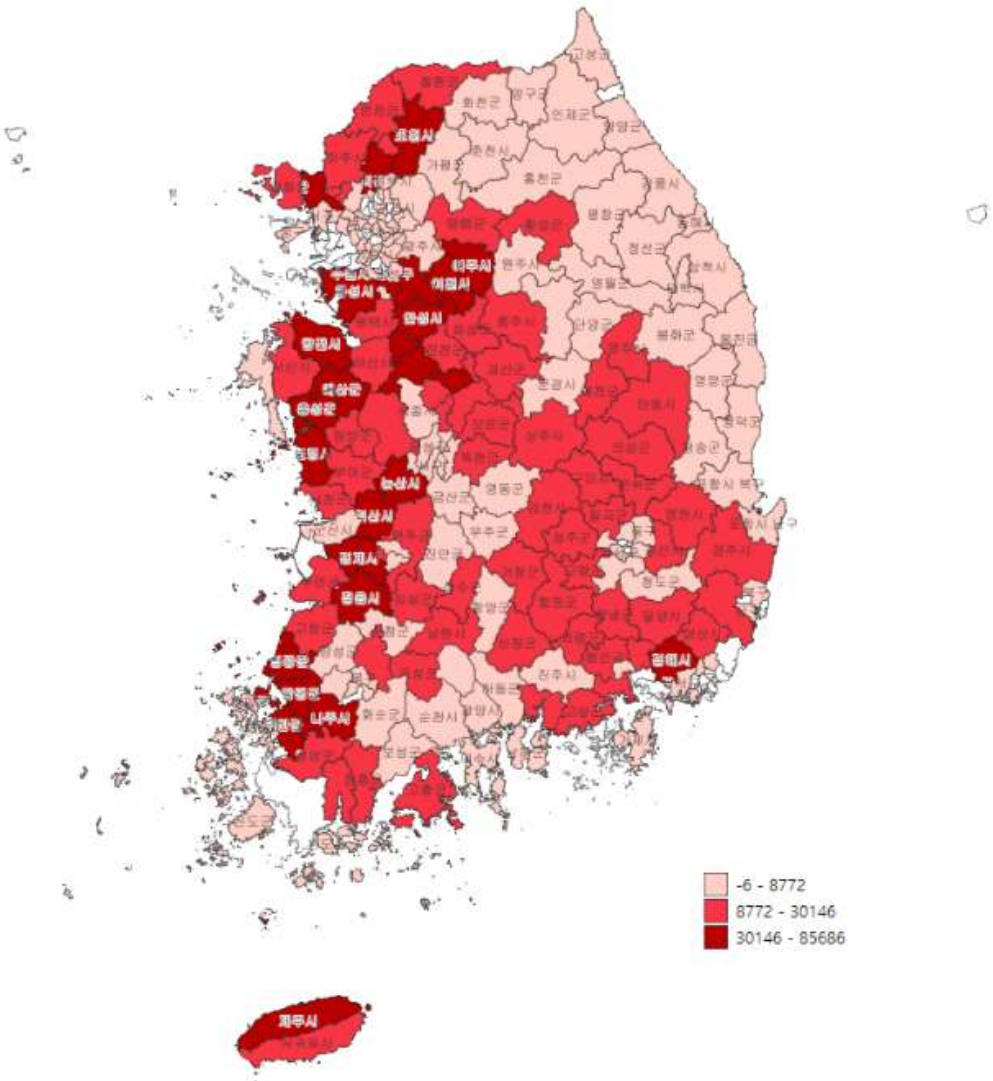
분석방법4. 환경오염도 지수 산출

```
new['index'] = new['시도별_가축더위지수'] * (-7.6110) + new['읍면동_총거주자수'] * 0.1391 + new['가축분뇨발생량_합계_면적비'] * 8.569e+04 + new['가축사육종사자수'] * (-0.0834) + new['분뇨처리업장_개수'] * 0.1123 + new['분뇨처리_정화'] * 0.0240
new.head(2)
```

	시도별_가축더위지수	읍면동_총거주자수	가축분뇨발생량_합계_면적비	가축사육종사자수	분뇨처리업장_개수	분뇨처리_정화	index
0	0.553841	0.574057	0.058740	0.721357	0.0	0.647682	5029.271643
1	0.553841	0.029032	0.026158	0.636634	0.0	0.187151	2237.224690

환경오염도 지수 산출식에 따라 회귀계수 가중치를 적용한 환경오염지수(index)를 개발하였다.

분석방법5. QGIS 지도 시각화



위험구역을 도출하기 위한 환경오염지수 구간을 설정하고, 해당 구간에 따라 QGIS에서 지도 시각화를 수행하였다.