

(재)미래와소프트웨어와 함께하는 제3회 아이디어 해커톤

# 빅데이터 활용 미래 사회문제 해결 아이디어 해커톤

인구 / 환경 / 생태계 데이터를 활용한 지역별  
독조 발생 원인 분석과 해결 방안 제시

팀명

G-MILE

팀원

고나경  
손유진

CONTENTS

목차

STEP 1.

분석 개요

주제 선정 배경 03

연구의 필요성 04

분석 프로세스 05

STEP 2.

분석 기획

데이터 수집 및 전처리 06-07

남조류 세포수 빈도 결정  
요인 파악 : 군집 분석

군집 분석 시각화

STEP 3.

분석 결과

남조류 세포수 발생 요인 파악

관계식 도출

STEP 4.

결론

해결 방안 제시

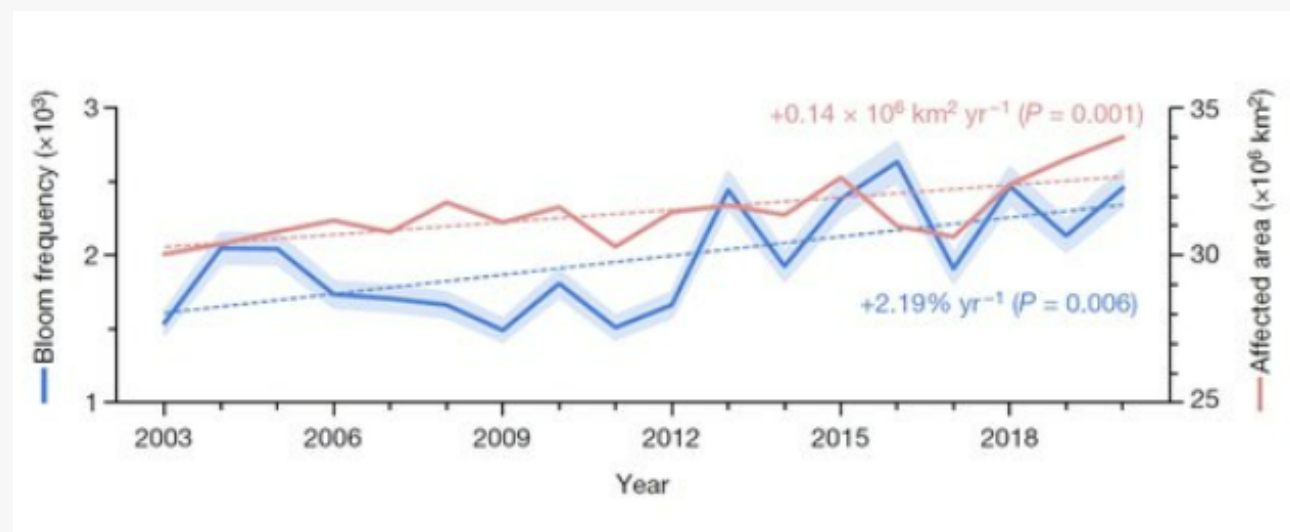
기대 효과 제시

데이터 목록과 참고문헌

# 주제 선정 배경

## [녹조 발생 빈도 증가]

- 중앙 일보에서 발표한 2002~2020년 사이 전 세계 해양에서 조류 대발생 해역의 면적과 관찰 빈도가 증가하였다고함.
- 아래 표는 전세계 해양의 조류 대발생추세, 파란색 선은 조류 대발생 빈도, 빨간색 선은 조류 대발생 피해면적을 나타냄



## [녹조 발생 원인과 녹조의 위험성]

- 한겨레 신문에 의하면, 녹조에서 나오는 독성물질로 액체로 마시거나 피부에 닿거나 호흡을 통해 사람의 몸에 흡수될 수 있으며, 간, 폐, 생식기, 신경계등에 악영향을 주는 발암물질이라 함
- 4대강 사업 이후 강물 체류 시간은 낙동강은 11.6배, 한강은 3.4배, 금강은 2.8배, 영산강은 7.7배 늘어났다고 함
- 이러한 녹조 발생의 주된 원인으로는 일사량, 수온, 유량, 유속 등을 원인으로 보고 있음.

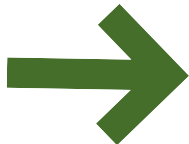


녹조의 위험성에 대처하기 위해서는, 기존의 발생 요인 이외의 원인을 파악한 후 녹조 발생에 대한 선제적 대응이 필요하다.

# 연구의 필요성

## [분산분석을 통하여 2014-2023년 서울시 9개의 대교의 남조류 세포수 비교 ]

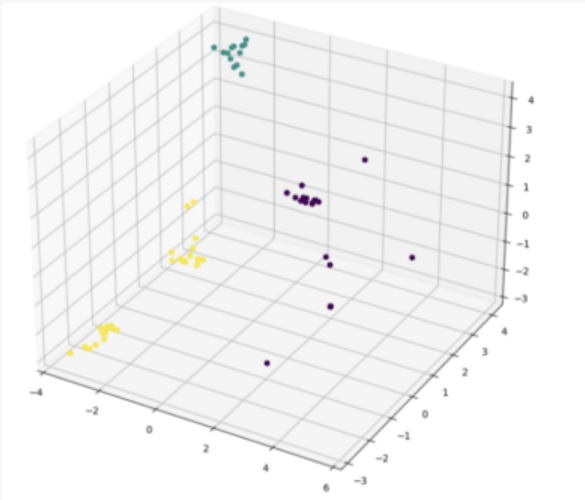
☑ 목 표	9개의 대교별 녹조류 세포수 관측값의 분포의 형태를 비교해보자							
☑ 가 설	H 0 : 대교(총 9가지의 대교) 에 따라 남조류 세포수에는 차이가 없을 것이다. H 1 : 하나의 대교에라도 남조류 세포수에는 차이가 존재할 것이다.							
	<p>p-unc</p> <p>2.943223e-14</p>	<p>→ P-value가 0.05 작으므로 H1 채택.</p> <p>따라서 어떤 대교와 어떤 대교 사이의 남조류 세포수의 차이가 존재하는지 확인 필요</p>						
☑ 사후검정	H 0 : 두개의 대교간의 남조류 세포수의 평균은 동일하다 H 1 : 두개의 대교간의 남조류 세포수의 평균은 동일하지 않다.							
	<table><thead><tr><th>A</th><th>B</th><th>pval</th></tr></thead><tbody><tr><td>강동대교</td><td>마포대교</td><td>0.004901</td></tr></tbody></table>	A	B	pval	강동대교	마포대교	0.004901	<p>P-value가 0.05 작으므로 H1 채택.</p> <p>두개의 대교간의 남조류 세포수가 다른 분포를 가짐</p> <p>총 36가지의 경우 중 34가지가 다른 분포를 가진다는 것을 확인</p>
A	B	pval						
강동대교	마포대교	0.004901						



9개의 대교 간 남조류 세포수 분포의 차이가 있다면, 이는 기존의 일사량, 수온, 유량등의 요인 이외에도 지역별 특성을 반영하는 것이다.

따라서, 녹조 발생 요인 을 파악한 후 , 지역별 남조류 세포수 차이의 원인을 분석하여 해결 방 안을 제시하려고 함

# 분석 프로세스

데이터 수집 및 전처리	남조류 세포수 발생 요인 파악	군집별 발생 요인 파악	분석 결과 및 기대효과
<div>서울 열린 데이터 광장</div> <div>공공데이터 포털</div> <div>물 환경 정보 시스템</div> <div>기상청</div> <div>서울특별시 TOPIS</div> <div>↓</div> <div>1<div>자연환경 데이터</div><div>수질, 기후, 녹조 등</div></div> <div>2<div>도시 및 인프라 데이터</div><div>교통량</div></div> <div>3<div>인구 및 사회 데이터</div><div>인구구조, 생활에너지 절약 등</div></div> <div>4<div>생태계 데이터</div><div>녹지대 위치정보</div></div>	<div>군집 분석</div> <div>k-prototype 클러스터링</div> <div></div> <div>↓<div>자세한 발생 요인 파악을 위하여, 변수의 개수 줄이기</div></div> <div>분산 분석</div> <div><div>• 등분산성 검정</div><div>• 분산 분석</div><div>• 사후검정</div></div> <div>&lt;변수의 개수&gt;</div> <div>44 → 22</div>	<div>1<div>후진제거법</div><div>다중공선성 제거하기 위해 중요한 변수 추출</div><div>&lt;변수의 개수&gt;</div><div>22 → 10</div></div> <div>2<div>RandomForestRegressor</div><div>변수 별 information 값을 통해 남조류 세포수 와 관련된 주요변수 추출</div></div> <div>3<div>다중회귀분석</div><div>추출된 주요 변수를 통해 회귀분석을 진행 한 후, 회귀 계수를 통하여 군집별 녹조 발 생 계산식 생성</div></div> <div>→ 군집별 남조류 세포수 발생요인 파악</div>	

# 데이터 수집 및 가공 : 2017년 - 2023년

## [자연 환경 데이터]

활용 데이터	전처리 과정	데이터 셋
물환경정보시스템 총량측정망	대교별 총량측정소의 위치를 확인하여, 가장 가까운 측정소에 매칭한 후, 녹조를 측정한 날의 값 사용	수소이온농도, 전기전도도, 용존산소, 부유물질, 총질소, 총인, 총유기탄소, 유량
물환경정보시스템 방사성물질측정망	대교별 방사선 물질 측정소의 위치를 확인하여, 가장 가까운 측정소에 매칭한 후, 녹조를 측정한 날의 값 사용	Cs-134(세슘),Cs-137(세슘),I-131(요오드)
물환경정보시스템 수질측정망	대교별 수질측정소의 위치를 확인하여, 가장 가까운 측정소에 매칭한 후, 녹조를 측정한 날의 값 사용	수온,DO,BOD,COD,TN,TP,TOC,SS
물재생센터 수질 현황	1. 서울시 4개 물재생센터의 위치를 확인해 대교를 가장 가까운 물재생센터에 매칭 2. [‘계통구분’] 컬럼의 값 중, 방류 값만 사용 3. 측정 일자별 각 물재생센터의 n개의 처리장의 값을 평균 내어 녹조를 측정한 날의 값 사용	'방류_BOD', '방류_TOC'(2021년 1월 이전자료는 COD)', '방류_SS', '방류_T-N', '방류_TP', '방류_COCG'
물재생센터 하수처리량	측정 일자별 각 물재생센터의 n개의 처리장의 값을 평균 내어 녹조를 측정한 날의 값 사용	1차 하수 처리장, 2차 하수처리장
서울특별시_한강대교별 녹조 관측정보	1. 2017-2023 중 9개의 대교중 서울특별시에만 속하는 7개의 대교만 사용 2. 지도상 위치를 파악하여, 대교별 맞닿은 상구/상동, 하구/하동의 지역을 구함	남조류 세포수 , 주소, 상구, 상동 , 하구, 하동
기상청 종관기상관측(ASOS)	1. 2017-2023 일 단위의 평균기온   평균 상대습도   일강수량 합계   일사량 네가지 자료를 수집 2. 일강수량(mm) - 0값으로 처리 3. 평균기온   평균 상대습도 - 선형보간법 사용	평균기온, 일강수량,평균 상대습도,합계 일사량
일강수량을 기준으로 파생변수 생성	달을 기준으로 하루에 80mm이상 비가 온 날의 횟수 계산	80mm이상비온날횟수

## [도시 및 인프라 데이터]

활용 데이터	전처리 과정	데이터 셋
서울시 지점별 일자별 교통량 정보	녹조를 관측한 날의 해당 대교별 각 교통량 사용	교통량

# 데이터 수집 및 가공 : 2017년 - 2023년

## [인구 및 사회 데이터]

활용 데이터	전처리 과정	데이터 셋
서울시 생활에너지 절약 정도 통계	1. 물재생센터가 관할하는 집계구를 경계로, 물재생센터를 기준으로 시민의식환경 지표를 연도별로 평균내어 활용 2. 결측값에 대하여 결측치 보간법을 사용하여 채움	시민의식환경
주민등록인구 데이터	1. 분기별로 조사된 주민등록인구 데이터를 결합하기 위하여, 녹조를 측정한 달이 1~2월이면 “, 3~5월이면 ‘1/4’, 6~8월이면 ‘1/2’, 9~12월이면 ‘3/4’의 값을 가지는 분기 컬럼을 생성해줌 2. 대교별 맞닿아있는 구와 동을 기준으로 녹조 관측값의 년도 / 분기_m에 해당하는 상구/상동, 하구/하동의 주민등록 인구데이터의 값을 사용 3. 마포동과 신천동의 데이터 부재로 인하여 각각 행정동인 용강동,도화동/잠실 4동,잠실 6동의 주민등록인구를 평균낸 값으로 대체	분기_m , 하구분기별인구수,상구분기별인구수,상동분기별인구수,하동분기별인구수
서울시 상권분석서비스(길단위인구-행정동)	1. 분기별로 조사된 주민등록인구 데이터를 결합하기 위하여, 녹조를 측정한 달이 1~2월이면 “, 3~5월이면 ‘1/4’, 6~8월이면 ‘1/2’, 9~12월이면 ‘3/4’의 값을 가지는 분기 컬럼을 생성해줌 2. 대교별 맞닿아있는 구와 동을 기준으로 녹조 관측값의 년도 / 분기_m에 해당하는 상구/상동, 하구/하동의 주민등록인구데이터의 값을 사용 3. 결측치를 fillna(method='ffill')방식으로 처리	'총_유동인구_수_상구','연령대_10 ~ 60 유동인구수_상구 (총 6개의 컬럼) 총_유동인구_수_하구','연령대_10 ~ 60 유동인구수_하구 (총 6개의 컬럼)

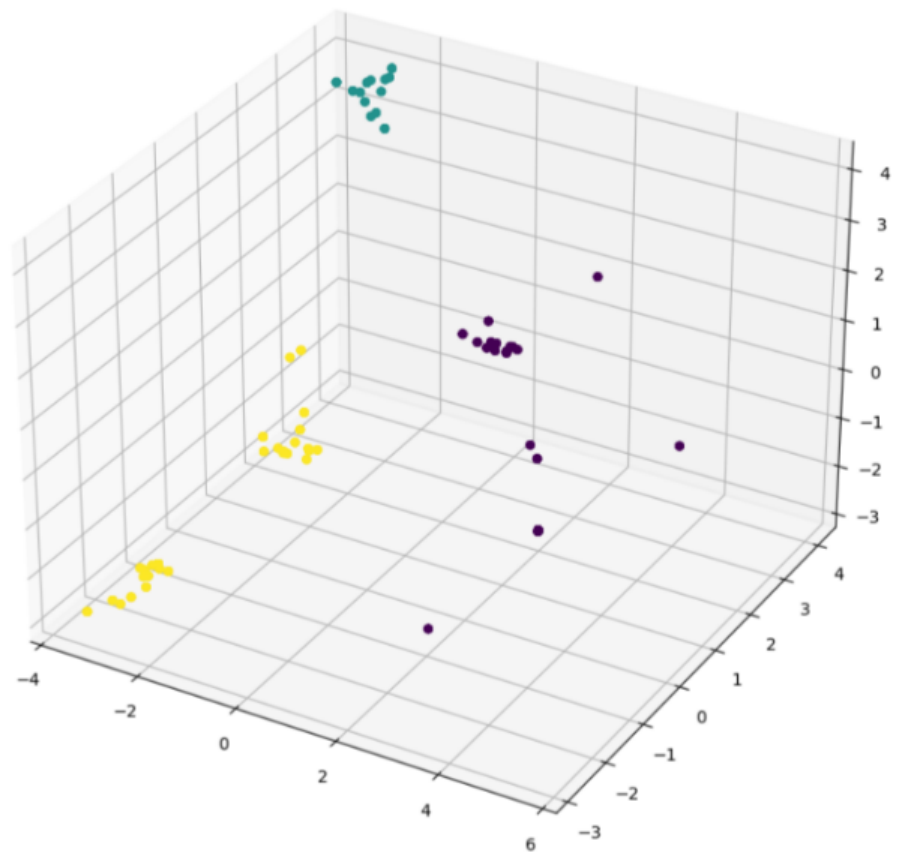
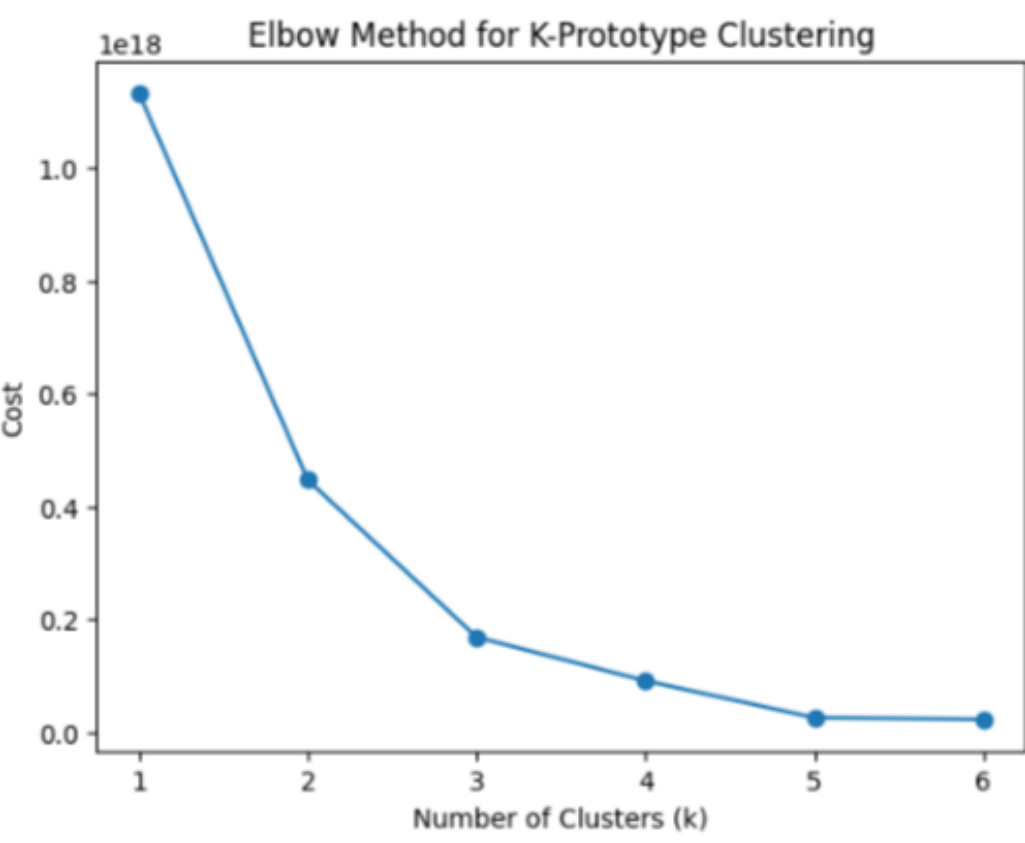
## [생태계 데이터 ]

활용 데이터	전처리 과정	데이터 셋
서울시 녹지대 위치정보	1. 녹지대분류 중 ‘하천변조경’,‘하천변’,‘하천변 녹지’의 값만 사용 2. 대교별 맞닿아있는 구를 기준으로 구별로 녹지대 면적의 합한 값 사용 ( 해당구에 할당된 녹지 면적이 없을 경우에는 0으로 처리)	'한강옆녹지_상구', '한강옆녹지_하구', '녹지_상구', '녹지_하구',

# 남조류 세포 수 빈도 결정 요인

- 군집화에 사용한 데이터프레임(968 rows × 65 columns)

	녹조	수소이온농도 (ph)	전기전도도 (μS/cm)	용존산소(mg/L)	BOD(mg/L)_x	COD(mg/L)_x	부유물질(mg/L)	총질소(T-N)(mg/L)	총인(T-P)(mg/L)	총유기탄소(TOC)(mg/L)	...	총_유동인구_수_상구	총_유동인구_수_하구	한강옆녹지_상구	한강옆녹지_하구	녹지_상구	녹지_하구	시민의식_환경
	mean	mean	mean	mean	mean	mean	mean	mean	mean	mean	...	mean	mean	mean	mean	mean	mean	mean
clusters																		
0	172.625304	8.166248	198.795777	9.324710	1.499717	4.168365	8.134786	2.895833	0.057142	2.670146	...	1.072918e+08	1.096986e+08	14.306569	9640.772482	1.295765e+06	1.434566e+06	6.075401
1	332.038186	8.187112	200.579356	10.212788	1.693099	4.518675	9.207578	2.954692	0.058849	2.611734	...	6.474160e+07	8.309658e+07	76.014320	2663.516945	6.041354e+05	1.205509e+06	5.687995
2	174.311594	8.318841	191.365942	9.562258	1.328925	4.043841	7.700362	2.346699	0.045803	2.497162	...	7.613340e+07	1.445553e+08	490.000000	0.000000	2.355850e+05	1.156247e+06	6.103913

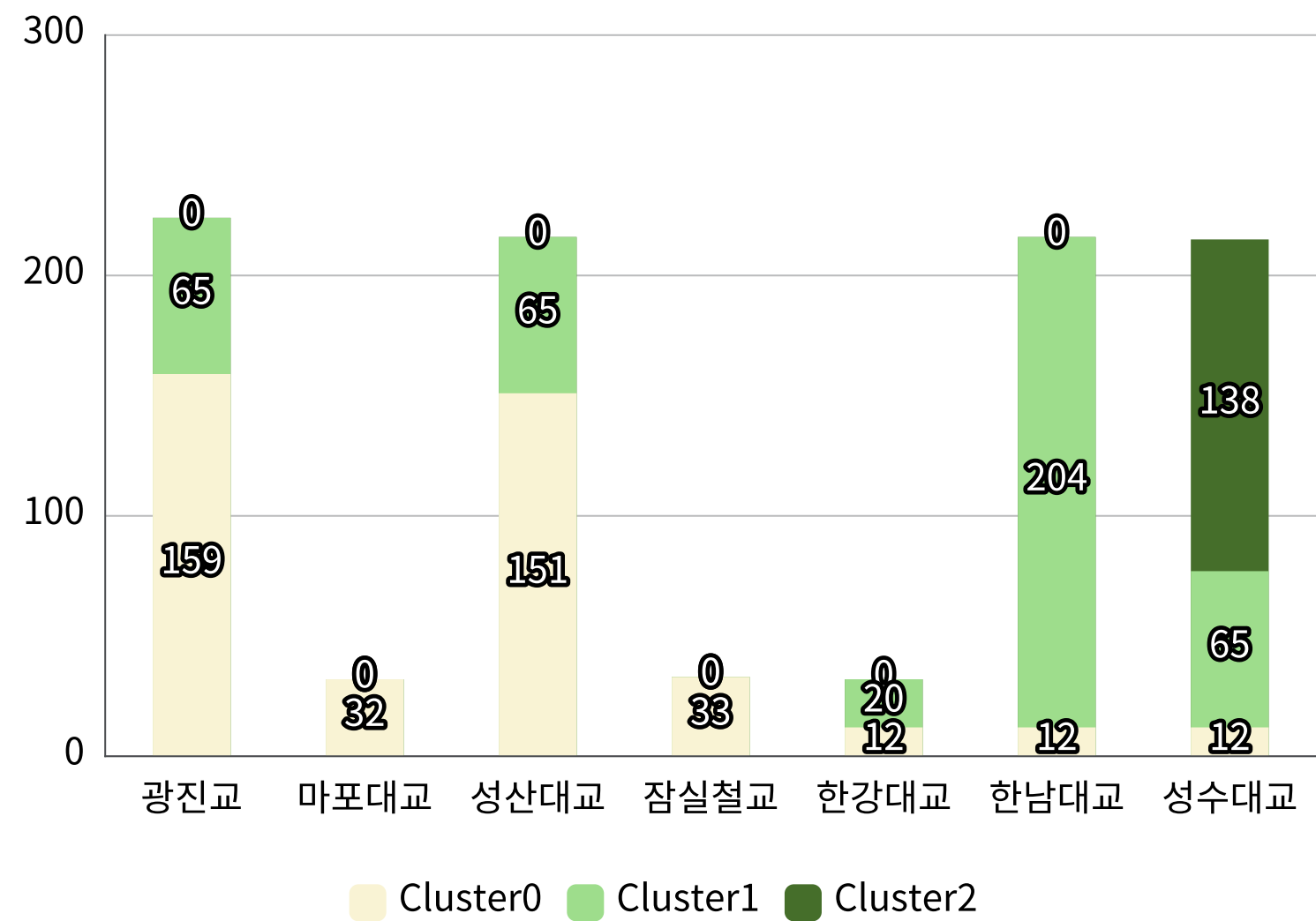


☑ 군집화 모델

k-Prototype(k=3)  
Elbow method 결과를 바탕으로 군집 수를 k=3으로 결정하였다.  
"대교에 따라 남조류 세포수에 차이가 있을 것이다."를 가설으로 하기에 범주형 변수인 "대교"를 함께 고려해주기 위해 k-Prototype 모델을 사용하였다.



# 군집 분석 결과



Cluster	count	mean
0	411	172.6
1	419	332
2	138	174.3

<군집별 수와 녹조 평균>

이후, 군집별 녹조의 원인을 밝히고  
해결방안을 도출하기 위해  
군집화 결과를 기반으로  
각 대교를 **최빈값 기준**  
하나의 군집에만 할당해주었다.  
Cluster0 : 광진교,마포대교,성산대교,  
잠실철교  
**Cluster1** : 한강대교,한남대교  
Cluster2 : 성수대교

# 군집 분석 결과 \_ 가설기반

## [ 가설 ]

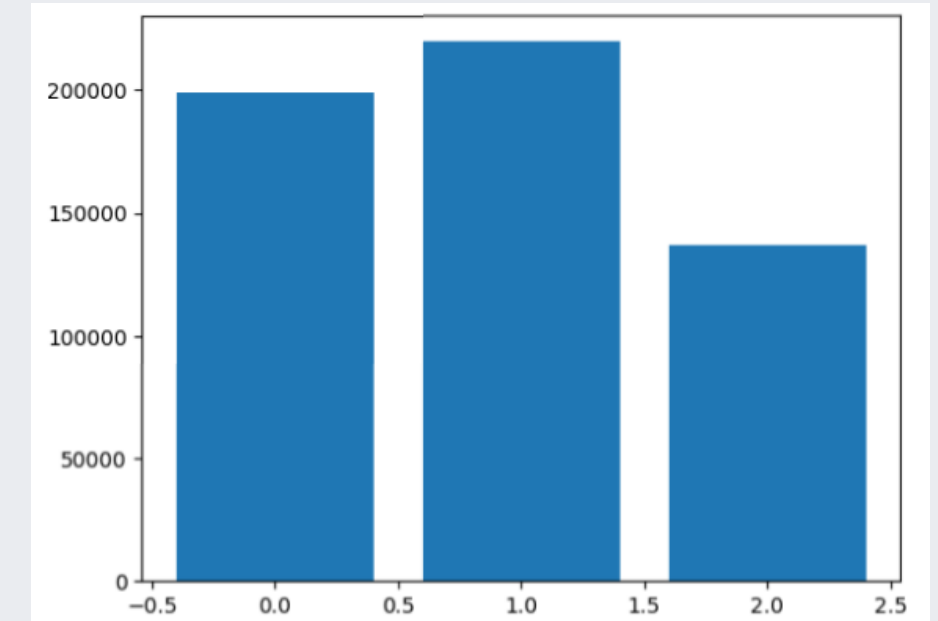
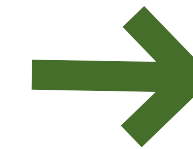
☑ 가설 1      환경에 대한 시민의식이 높으면 녹조가 적을 것이다.

☑ 가설 2      대교 주변 녹지가 많으면 녹조가 적을 것이다.

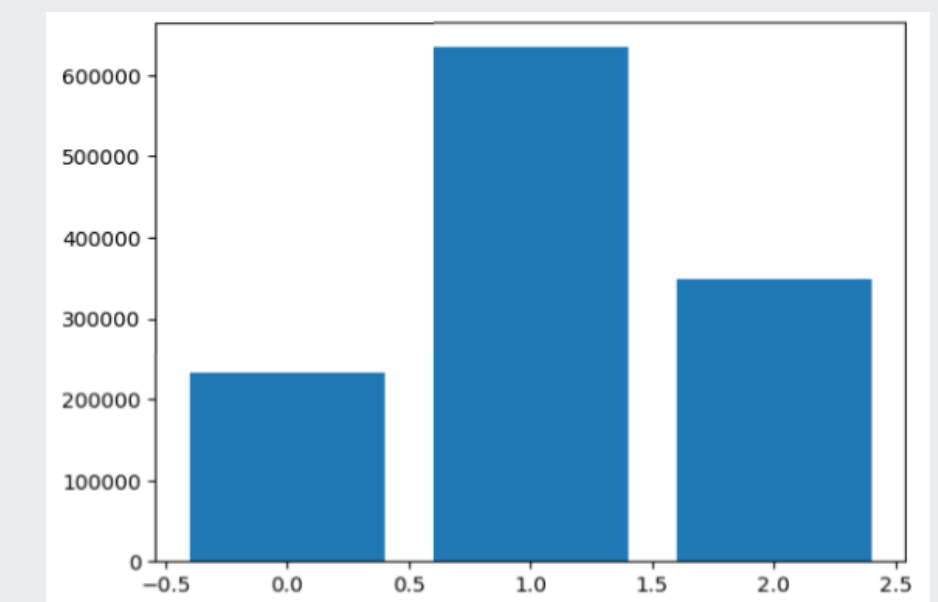
☑ 가설 3      교통량이 많으면 녹조가 많을 것이다.

☑ 가설 4      방출되는 하수의 수질이 안 좋으면 해당되는 물재생센터 주변 대교에 녹조가 많을 것이다.

☑ 가설 5      하수처리량이 많으면 녹조가 많을 것이다.



<군집별 해당하는 대교에서의 교통량>



<군집별 해당하는 물재생센터에서의 1차하수처리량>

# 군집별 분산분석

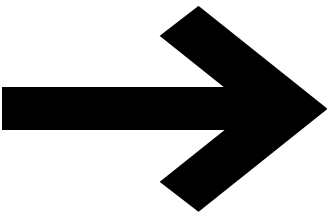
총 n개의 데이터 중 유의미한 변수 추출의 필요성

☑ 목 표	3개의 군집별 변수의 분포의 형태를 비교해보자
☑ 가 설	H 0 : 군집 (총 3개) 에 따라 각 변수의 값에는 차이가 없을 것이다. H 1 : 하나의 군집에라도 각 변수의 값에는 차이가 존재할 것이다.
↓ 사후 검정이 필요한 변수 총 38개	
☑ 사후검정	H 0 : 두개의 군집간 변수의 값의 평균은 동일하다 H 1 : 두개의 군집간 변수의 값의 평균은 동일하지 않다.

↓ <총질소(T-N)(mg/L)에 대하여 사후 검정 예시>

A	B	pval
0	1	0.6436029
0	2	0.0000000
1	2	0.0000000

총 3가지의 경우 중 2가지의 pval가 0.05보다 작다면 (다른 분포를 가진다면) 이는 군집별 유의미한 변수로 채택함



<최종 변수의 개수>

44

→

22

# 군집별 남조류 세포수 발생 요인 파악

군집화를 통해 나온 군집의 특성 - 군집 0

1

후진제거법

<변수의 개수>

22

→

10

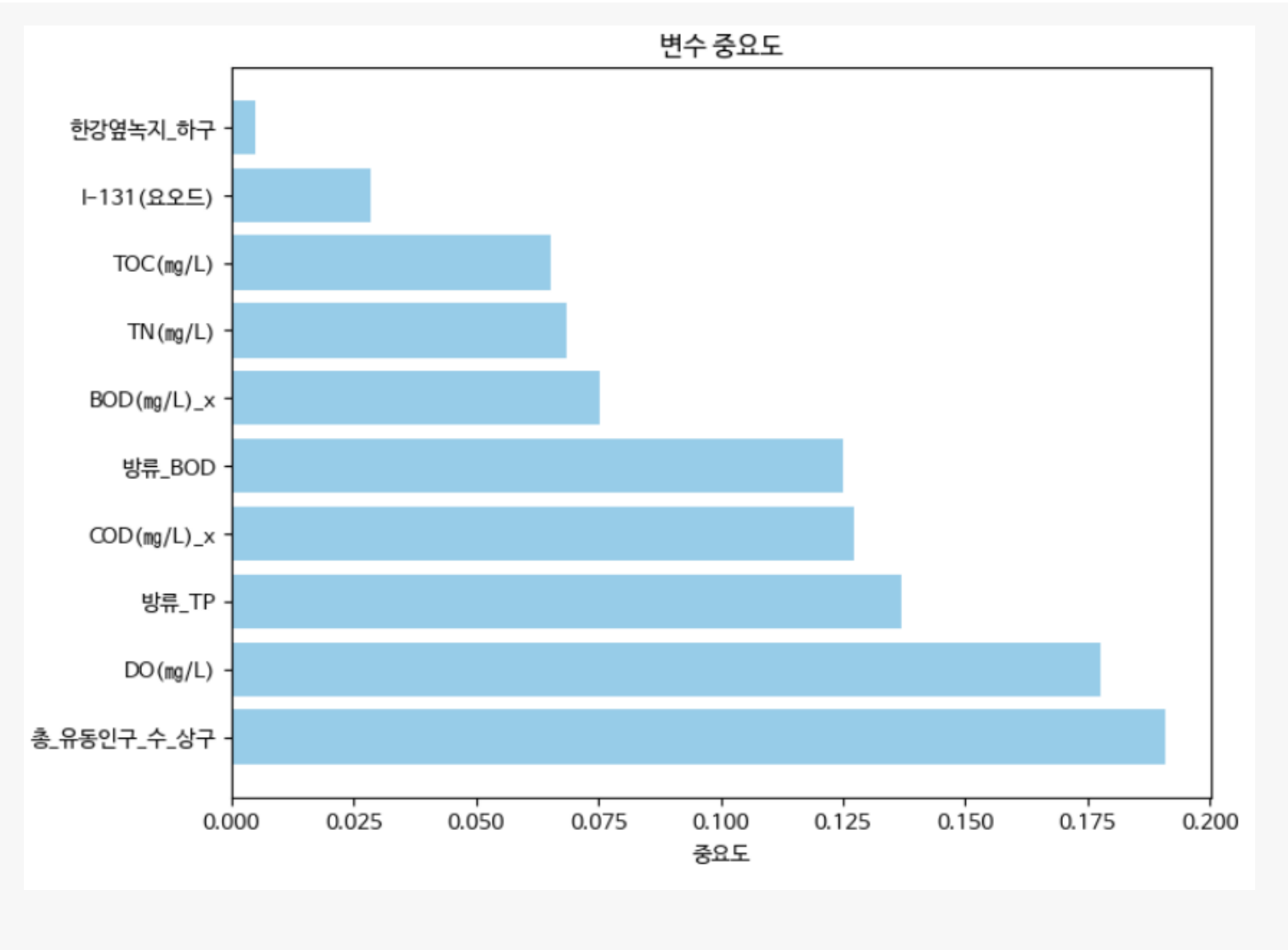
<변수 종류>

자연 환경 데이터	BOD(mg/L) COD(mg/L) I-131(요오드) DO(mg/L) TN(mg/L) TOC(mg/L) 방류_BOD 방류_TP
인구 및 사회 데이터	총_유동인구_수_상구
생태계 데이터	한강옆녹지_하구

2

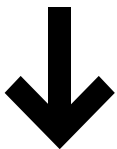
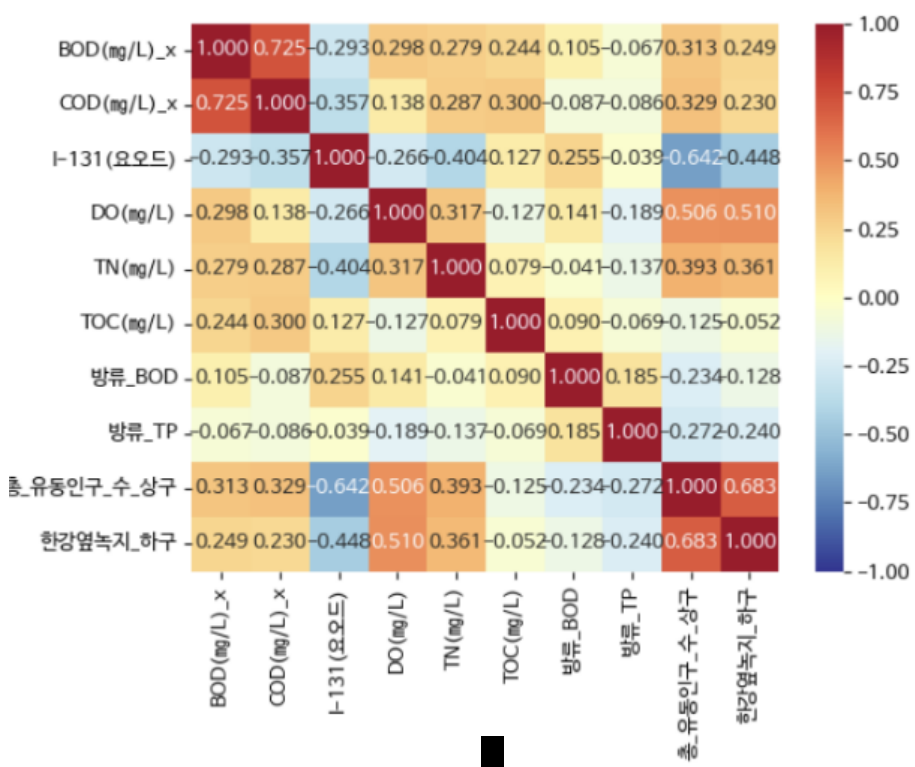
변수 중요도

후진 제거법을 통해 결정된 10개의 변수로, 남조류 세포수 예측을 통한  
변수 중요도 값 구함



3

요인 파악



군집 0의 남조류 세포수가 발생하는 주된  
요인은 수질 데이터이다.

또한 수질 데이터와 상구의 유동인구 수  
사이에는 상관관계가 높은 것으로 보여,  
개인의 노력을 통한 실질적인 수질 개선  
및 잦은 모니터링이 필요할 것으로  
여겨진다.

# 군집별 남조류 세포수 발생 요인 파악

## 군집화를 통해 나온 군집의 특성 - 군집 1

1

후진제거법

<변수의 개수>

25

→

10

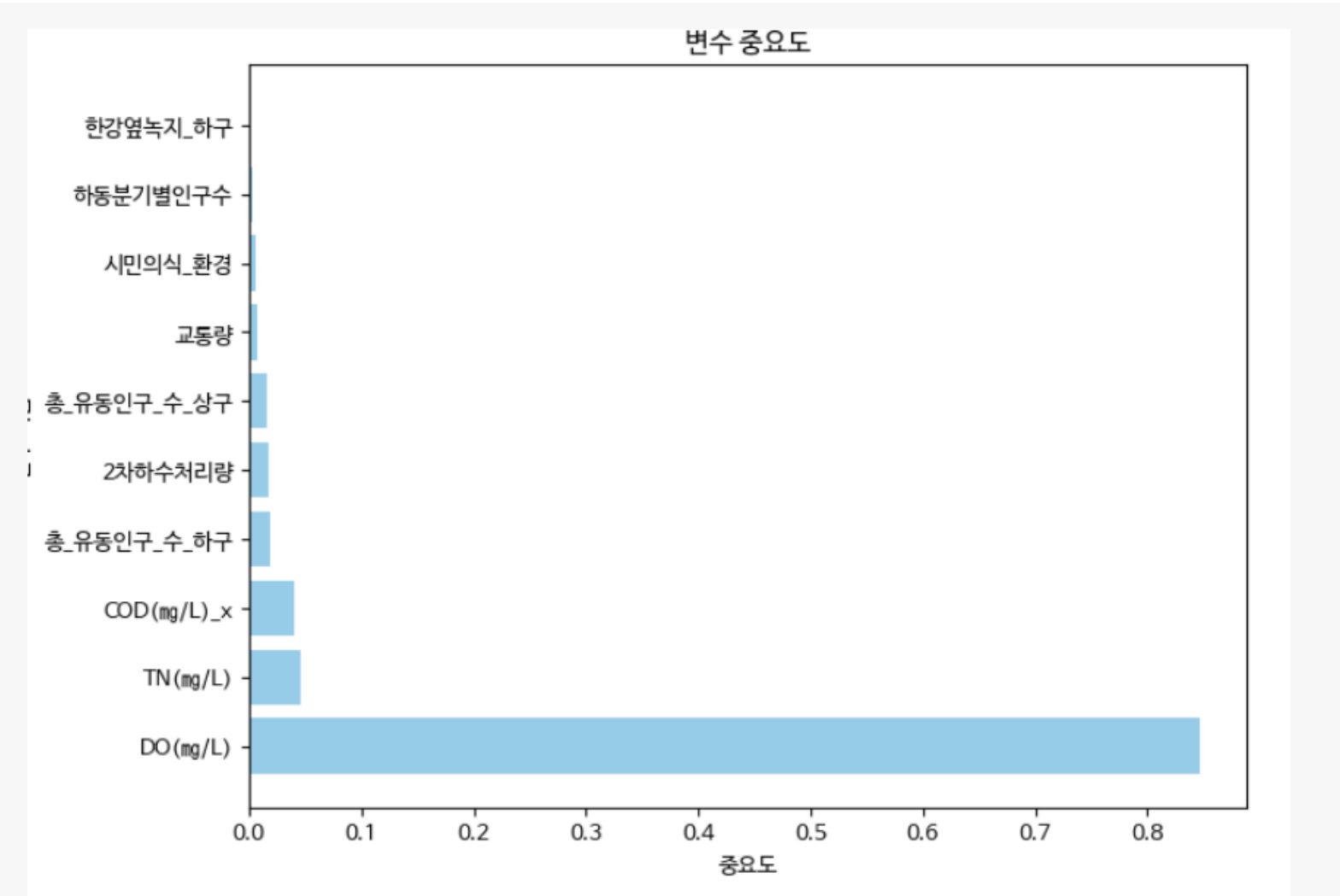
<변수 종류>

자연 환경 데이터	COD(mg/L) DO(mg/L) TN(mg/L) 2차하수처리량
인구 및 사회 데이터	총_유동인구_수_상구 총_유동인구_수_하구 하동분기별인구수 시민의식_환경
도시 및 인프라 데이터	교통량
생태계 데이터	한강옆녹지_하구

2

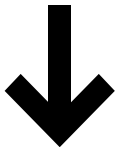
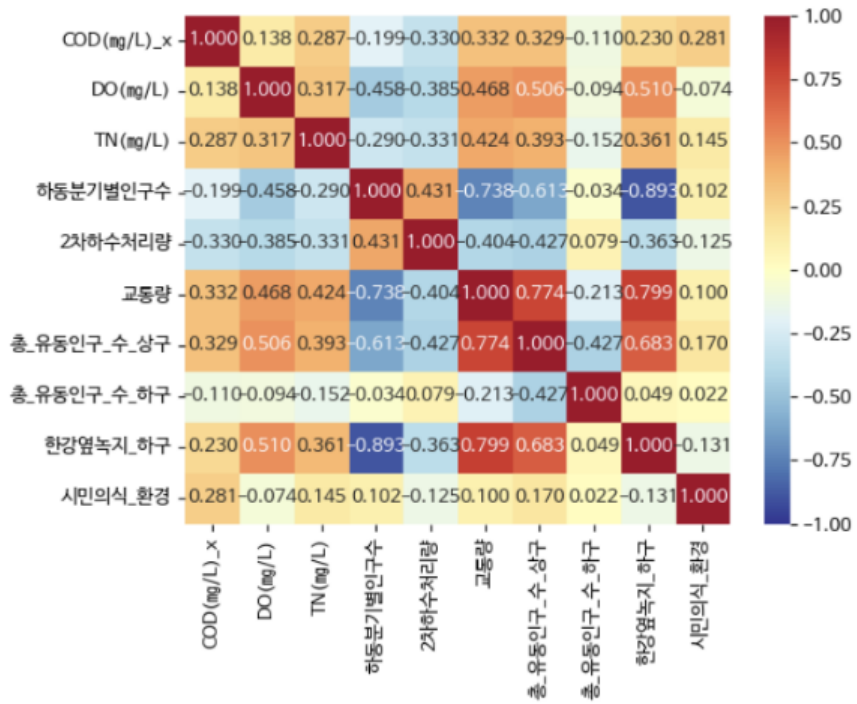
변수 중요도

후진 제거법을 통해 결정된 10개의 변수로, 남조류 세포수 예측을 통한  
변수 중요도 값 구함



3

요인 파악



군집 1의 남조류 세포수가 발생하는 주된  
요인은 인구수 / 교통량 데이터 때문이다.

또한 상구의 유동인구 수는 다른 주요 변수  
들과 모두 상관관계가 높은 것을 보여져, 다  
른 군집에 비해 개인의 활동에 높은 초점을  
맞추는 것이 중요하다고 여겨진다.

# 군집별 남조류 세포수 발생 요인 파악

## 군집화를 통해 나온 군집의 특성 - 군집 2

1

후진제거법

<변수의 개수>

25

→

10

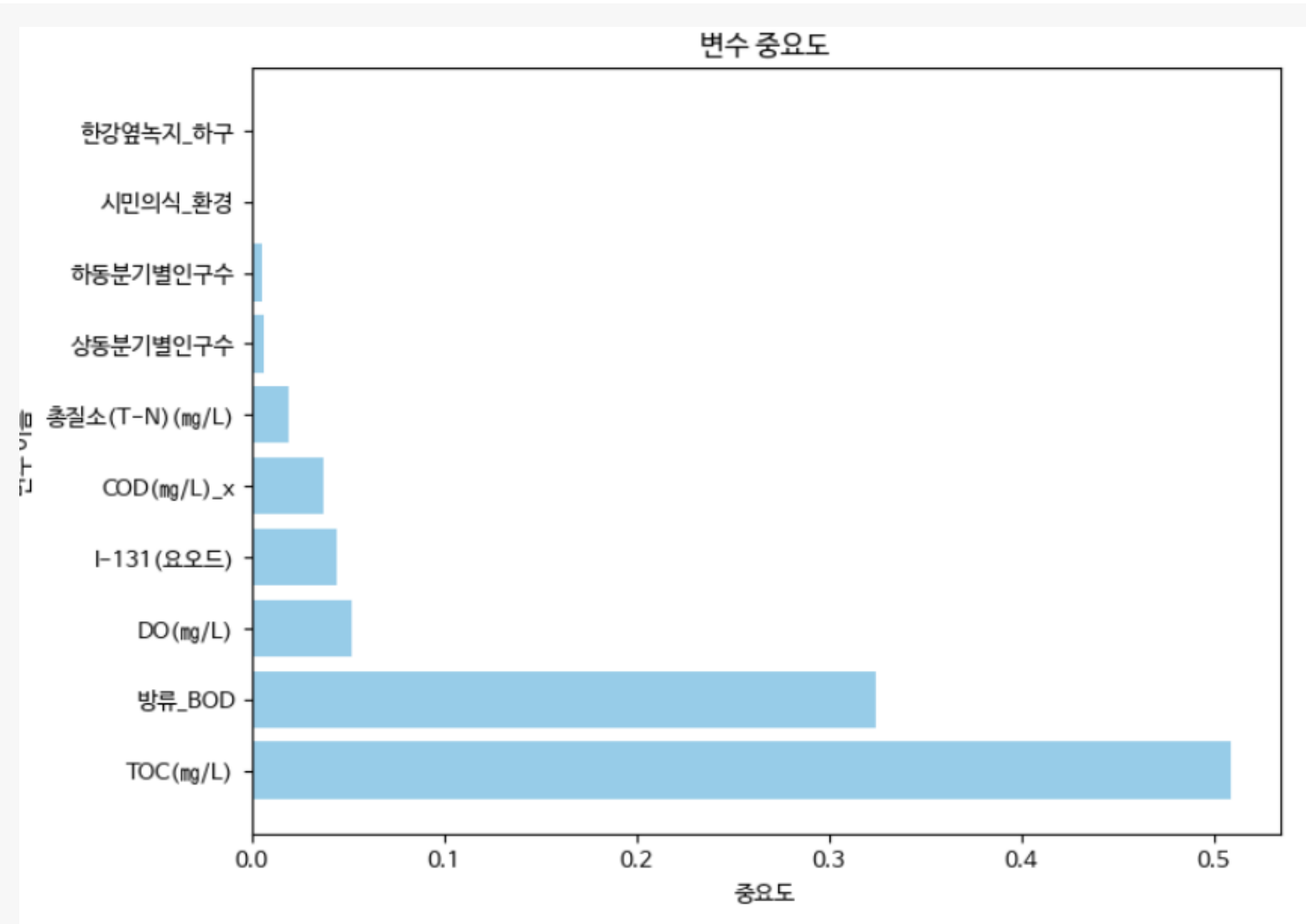
<변수 종류>

자연 환경 데이터	COD(mg/L) 총질소(T-N)(mg/L) I-131(요오드) DO(mg/L) TOC(mg/L) 방류_BOD
인구 및 사회 데이터	하동분기별인구수 상동분기별인구수 시민의식_환경
생태계 데이터	한강옆녹지_하구

2

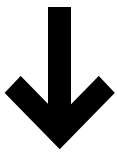
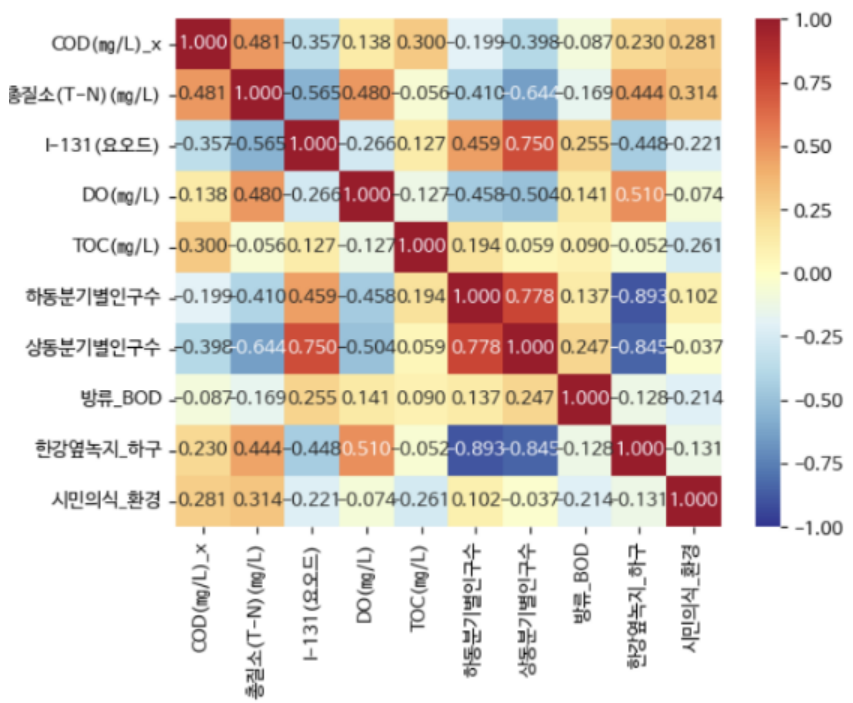
변수 중요도

후진 제거법을 통해 결정된 10개의 변수로, 남조류 세포수 예측을 통한  
변수 중요도 값 구함



3

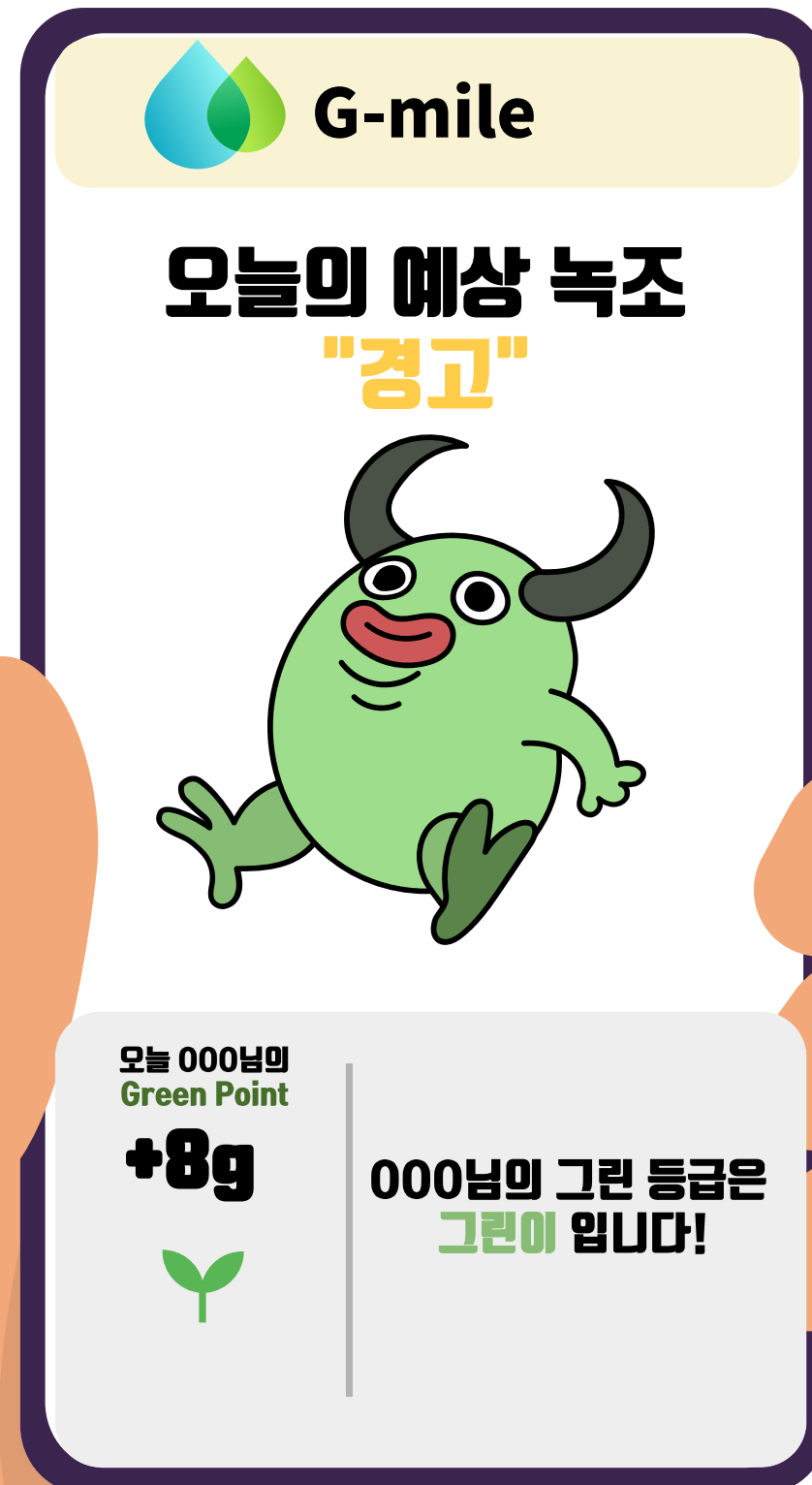
요인 파악



군집 2는 다른 군집에 비해, 상동/하동 분기  
별 인구수 즉 녹조 측정 지역 근처에 사는 거  
주민의 영향을 받는 것으로 보여진다.

또한 한강 옆 녹지 하구의 값이 다른 변수들  
이 상관관계가 높은 것으로 보여져, 녹지의  
주기적인 관리와 주변 거주민의 협조가 필  
요할 것으로 보인다.

# 해결방안 : G-mile



## G-mile 기능

- **탄소 예측 :**  
첫 화면에서 오늘의 탄소 상황을 확인할 수 있다.
- **그린 마일리지 적립 :**  
개인/거주지를 기준으로 팀을 나누어 그린 마일리지를 적립할 수 있다.  
매일 해당 지역을 기준으로 미션이 주어지고 해당 미션 수행 시, 그린 마일리지를 획득한다.  
특히, 개인단위로 적립된 마일리지에 대해서는 **그린 등급**이 주어진다.

→ 해당 등급은 정부 혹은 ESG경영에 관심있는 회사에서 시민의 **그린 마일리지 등급에 따른 혜택을** 줌으로서 환경 보호에 참여하면서 홍보효과를 얻고, 시민의 녹지 예방 활동을 촉진시킬 수 있다.





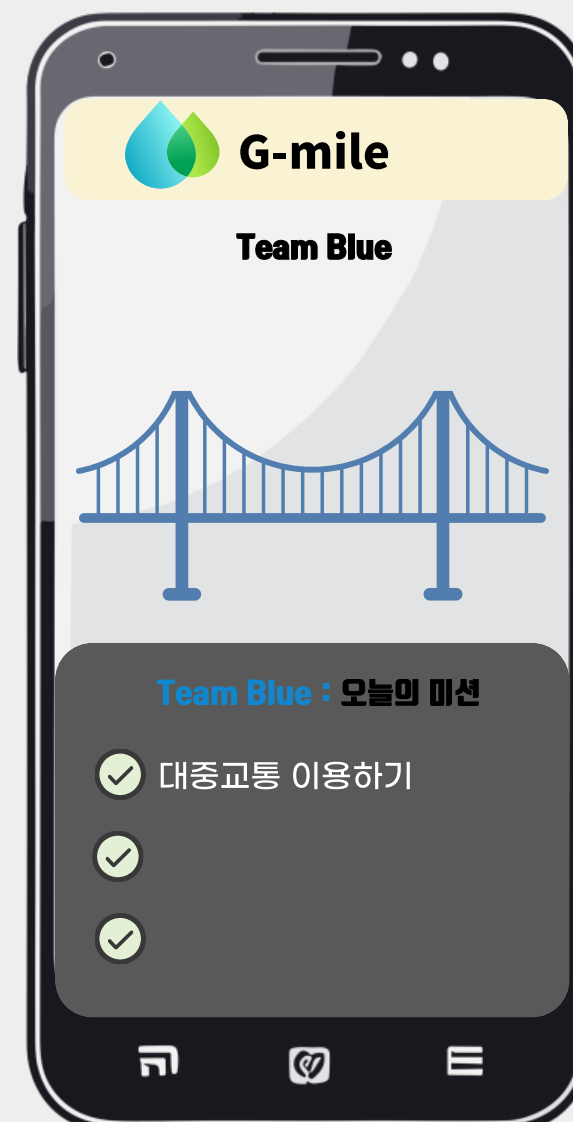
# 해결방안 : G-mile

- 각 대교가 해당되는 군집별로 나온 주요원인에 최적화된 시민 단합 하루미션 : G-mile

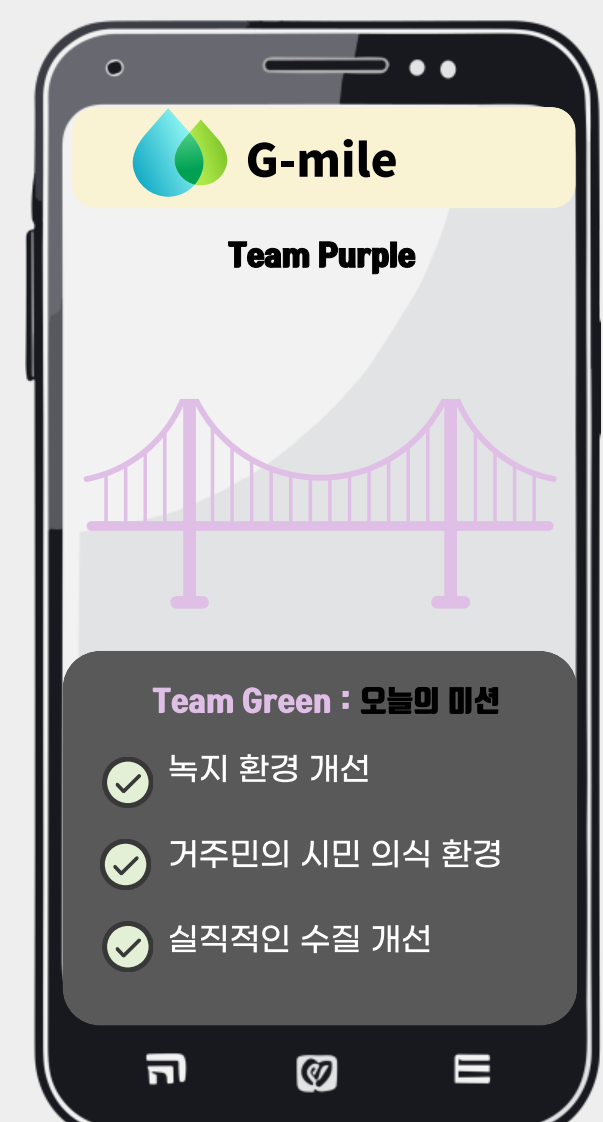
군집 0  
광진교, 마포대교, 성산대교, 잠실철교, 한남대교



군집1  
한강대교, 한남대교



군집2  
성수대교





# 기대효과

- ① 녹조에 환경에 대한 시민의 관심도가 분명한 영향을 미침을 규명하였다.  
녹조 외에도 다양한 환경 문제에서 시민 의식의 유의미한 영향을 발견한다면  
환경에 대한 단순한 교육/캠페인만으로 다양한 환경문제를 사전예방할 수 있을 것이다.
- ② 단순 정부 차원에서의 해결이 아닌 시민의 참여를 유도/녹조 정보 제공한다.
- ③ 각 군집에 해당되는 대교별로 녹조 발생 원인을 도출했기에 해당 원인에 집중한다면  
관리 대비 높은 효율을 낼 수 있을 것이다.