환경이 좋은 지역은 주택 공실률이 낮을까?

빅데이터 분석 기말 프로젝트

소프트웨어융합학과 2020111406 이해현(팀장)
소프트웨어융합학과 2020110381 오남의(팀원)

CONTENTS

1. 프로젝트 기획 배경

프로젝트 기획 배경

환경이 좋은 지역은 정말 주택의 공실률이 낮을까?

서울시 주택지역 중 공실률이 높은 지역의 이유가 해당 지역의 환경문제와
관련이 되어있는지 분석해보자!

주거지 선정에 있어 가장 많이 영향을 끼칠 수 있는 환경오염 지표로 **대기오염, 소음오염, 수질오염**을 선정

서울시 주택 공실률 현황 데이터	서울시 소음진동 민원 현황 데이터	서울시 대기오염발생 현황 데이터	서울시 환경오염물질 배출시설 현황 데이터

환경 오염과 서울시 주택 공실률의
상관관계 분석

주요 환경오염 요인

대기 오염

미세먼지

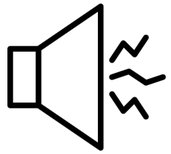
- 호흡기 및 심혈관 질환의 원인

질소산화물 및 황산화물

- 교통량이 많은 지역에서 주로 발생, 산성비의 원인

오존

- 높은 수준의 오존 노출은 호흡기 문제와 같은 건강문제 유발



소음 오염

교통 소음

- 도로, 철도, 공항 근처에 거주할 경우, 소음이 상시적인 스트레스 요인으로 작용

산업 소음

- 공장이나 산업 시설 근처에 거주할 경우, 높은 수준의 소음에 노출됨

수질 오염

산업 폐수

- 공장이나 처리 시설에서 나오는 폐수로 인한 지역 수질 오염

농약 및 비료

- 농촌 지역에서 사용되는 농약, 비료가 지하수 등 인근 수역 오염 가능

2. 프로젝트 가설 및 목표

프로젝트 가설 및 목표

프로젝트 가설

서울시 시민들은 주택 선택시 거주지역의 환경오염(대기·소음·수질)을 고려해서 선택 할 것이다

→ 환경오염(대기·소음·수질)이 적은 행정구의 주택 공실률이 낮을 것이다

프로젝트 목표

1) 서울시 주택공실률 현황, 소음·진동 민원, 대기오염, 환경오염물질 배출시설 데이터 수집 및 통합

2) 머신러닝 기반의 통계적 분석을 수행해 서울시 행정구 별 환경오염 수준과 주택 공실률 사이의 상관관계 규명

3. 데이터 수집 및 전처리

최근 3개년 (2022, 2021, 2020) 데이터 수집

데이터 수집 및 전처리 : 데이터 소개

서울시 주택 공실률 데이터 - 서울시 빈집 현황(구별) 통계

년도	행정구별	소계	단독주택	아파트	연립주택	다세대 주택	비거주용 건물내 주택
2020	종로구	2194	330	566	310	896	92

대기오염 - 서울시 기간별 일평균 대기환경 정보

년도	행정구별	이산화질소 농도(ppm)	오존농도 (ppm)	아황산가스 농도(ppm)	일산화가스 농도(ppm)	미세먼지 농도(㎍/㎥)	초미세먼지 농도(㎍/㎥)
2020	강남구	0.037	0.003	0.003	0.7	31.0	23.0

소음오염 - 서울시 소음진동민원 현황 통계

년도	행정구별	소음_공장	소음_교통	소음_생활	소음_소계
2020	강남구	0	0	7518	7518

대기, 수질오염 - 서울시 환경오염물질 배출 시설 통계

년도	행정구별	대기 1종	대기 2종	대기 3종	대기 4종	대기 5종	대기 소계	수질 1종	수질 2종	수질 3종	수질 4종	수질 5종	수질 소계
2020	강남구	4	1	2	57	122	186	0	2	6	7	167	182

데이터 수집 및 전처리 : 데이터 전처리

서울시 주택 공실률 데이터 - 서울시 빈집 현황(구별) 통계

- 그대로 사용

대기오염 - 서울시 기간별 일평균 대기환경 정보

- null 처리 : Null 값이 있는 경우 해당 데이터의 행정구 별 평균값으로 보간
- 연도 칼럼만 사용 (일자 버리기)

소음오염 - 서울시 소음진동민원 현황 통계

- 모든 값 int 로 데이터 타입 통일
- null 처리 : 공장, 교통 부분에 있는 (-) → 0
- 행정구별(1) ,소계 행 제거
- 연도칼럼 추가 (2022, 2021, 2020)

대기, 수질오염 - 서울시 환경오염물질 배출시설 통계

- 모든 값 int 로 데이터 타입 통일
- null 처리 : 공장, 교통 부분에 있는 (-) → 0
- 지역별(1) 제거, 소계 행 버리기
- 대기, 수질 부분 컬럼명 통일
- 소음및진동 열 제거
- 연도칼럼 추가 (2022, 2021, 2020)

최종 데이터 셋 및 환경오염 요인 간단 설명

	year	gu	빈집_소계	단독주택	아파트	연립주택	다세대주택	비거주용_건물내_주택	이산화질소농도(ppm)	오존농도(ppm)	일산화탄소농도(ppm)	아황산가스농도(ppm)	미세먼지농도(㎍/㎥)	초미세먼지농도(㎍/㎥)	소음_소계	대기_소계	수질_소계
0	2020	종로구	2194	330	566	310	896	92	0.041	0.003	0.7	0.003	35.0	25.0	2016	111	206
1	2020	종로구	2194	330	566	310	896	92	0.045	0.002	0.9	0.004	49.0	37.0	2016	111	206
2	2020	종로구	2194	330	566	310	896	92	0.038	0.010	0.9	0.004	66.0	48.0	2016	111	206
3	2020	종로구	2194	330	566	310	896	92	0.042	0.013	1.0	0.005	55.0	41.0	2016	111	206
4	2020	종로구	2194	330	566	310	896	92	0.050	0.007	1.0	0.005	54.0	40.0	2016	111	206

대기오염 요인

이산화질소농도(ppm), 오존 농도(ppm),
일산화탄소농도(ppm), 아황산가스농도(ppm),
미세먼지농도(㎍/㎥), 초미세먼지농도(㎍/㎥)

영향

호흡기 질환, 심혈관 질환, 폐 기능 저하, 두통,
어지러움, 암 발병 위험

환경오염물질 배출시설 (수질·대기)

오염물질 배출량: **1종 > 2종 > 3종 > 4종 > 5종**

1종: 가장 큰 규모의 매우 많은 오염물질 배출

5종: 가장 작은 규모의 오염물질 배출

4. 데이터 분석

1. 상관계수 기반의 통계적 EDA

상관계수(correlation coefficient)

변수 간에 존재하는 **선형 관계의 강도와 방향**을 나타내는 값

1 : 완전한 **양의 선형 관계** (두 변수는 동일한 비율로 증가)

-1 : 완전한 **음의 선형 관계** (한 변수가 증가할 때 다른 변수는 동일한 비율로 감소)

0 : 선형 관계가 없음 (두 변수 간 선형적인 연관이 없음)

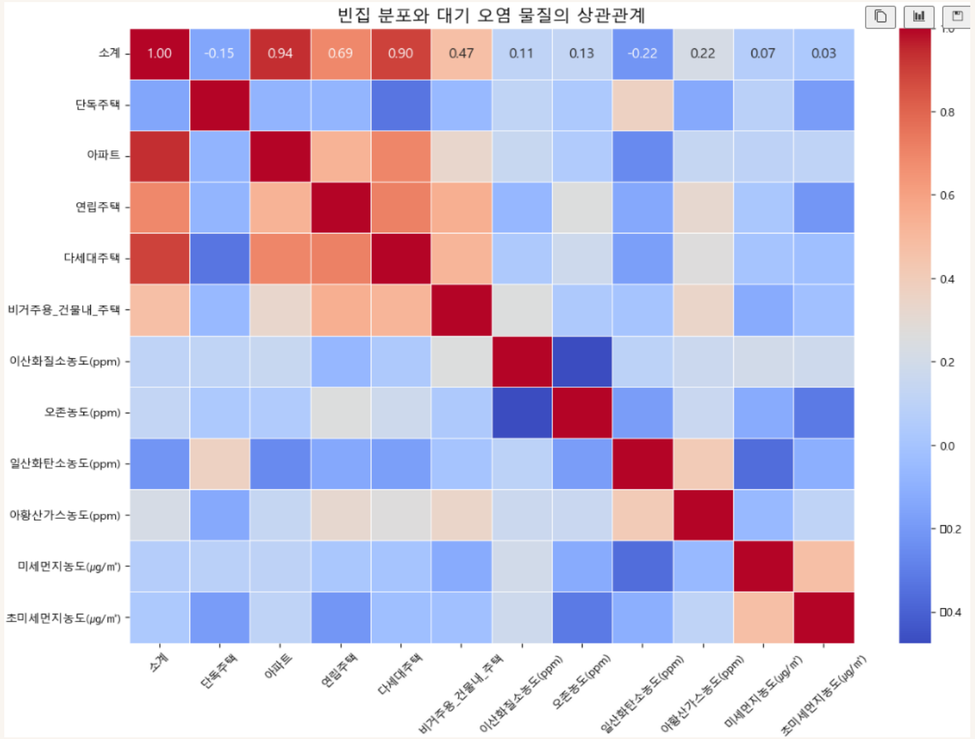
↓

주택 공실률 데이터와

각 환경오염 데이터(대기오염물질발생률, 소음오염민원발생률, 환경오염물질 배출시설 개수)의

상관행렬을 히트맵으로 **시각화**

서울시 주택공실률과 대기오염 데이터의 상관관계 분석



상관관계가 가장 높은 변수 조합

주택변수	대기오염 변수	|상관계수|	관계 종류
단독주택	일산화탄소농도(ppm)	0.3719	(-)
비거주용 건물 내 주택	아황산가스농도(ppm)	0.3422	(+)
연립주택	아황산가스농도(ppm)	0.3178	(+)
다세대주택	아황산가스농도(ppm)	0.2649	(+)
비거주용 건물 내 주택	이산화질소농도(ppm)	0.2579	(+)
소계	일산화탄소농도(ppm)	0.2575	(-)
소계	아황산가스농도(ppm)	0.2178	(-)

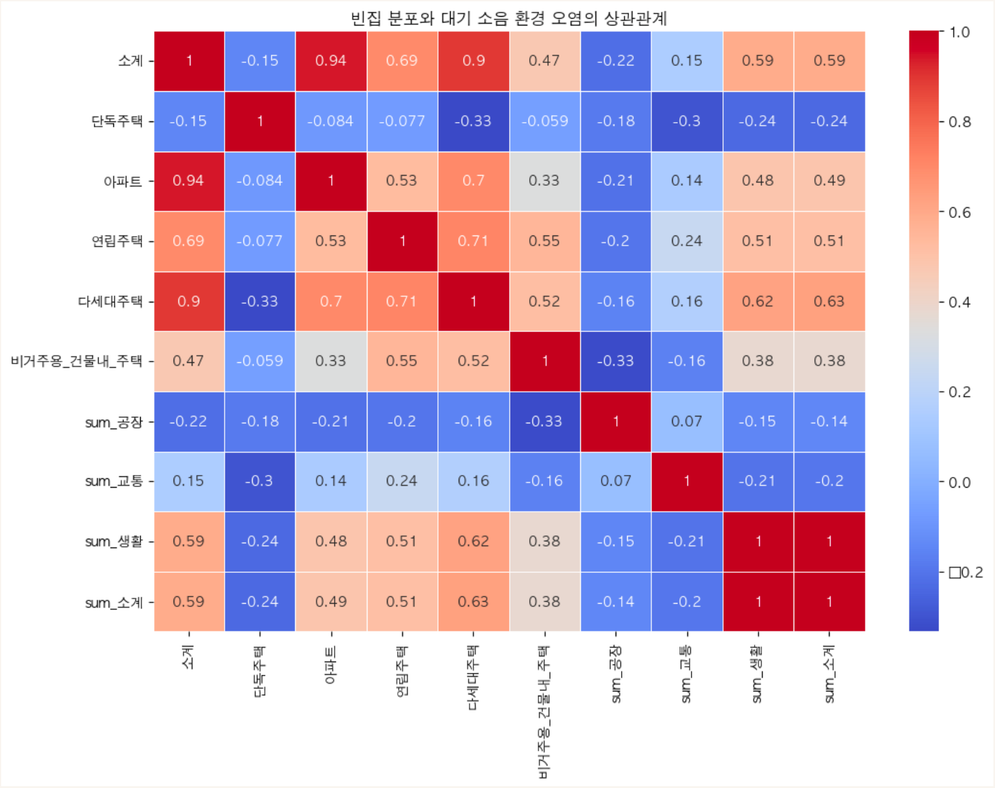
‘빈집_소계’ 와 ‘일산화탄소농도(ppm)’, ‘아황산가스농도(ppm)’ 사이의 **음의 상관관계** 확인

↓

공실률이 높은 지역일수록 일산화탄소와 아황산가스의 농도가 낮음

주거 밀집도가 낮아 차량 통행량이 적고, 산업 활동이 적어 대기오염물질 농도가 낮을 수 있음

서울시 주택공실률과 소음발생민원 데이터의 상관관계 분석



상관관계가 가장 높은 변수 조합

주택변수	대기오염 변수	|상관계수|	관계 종류
다세대주택	소음_소계	0.6262	(+)
다세대주택	소음_생활	0.6242	(+)
소계	소음_소계	0.5917	(+)
소계	소음_생활	0.5902	(+)
연립주택	소음_소계	0.5128	(+)

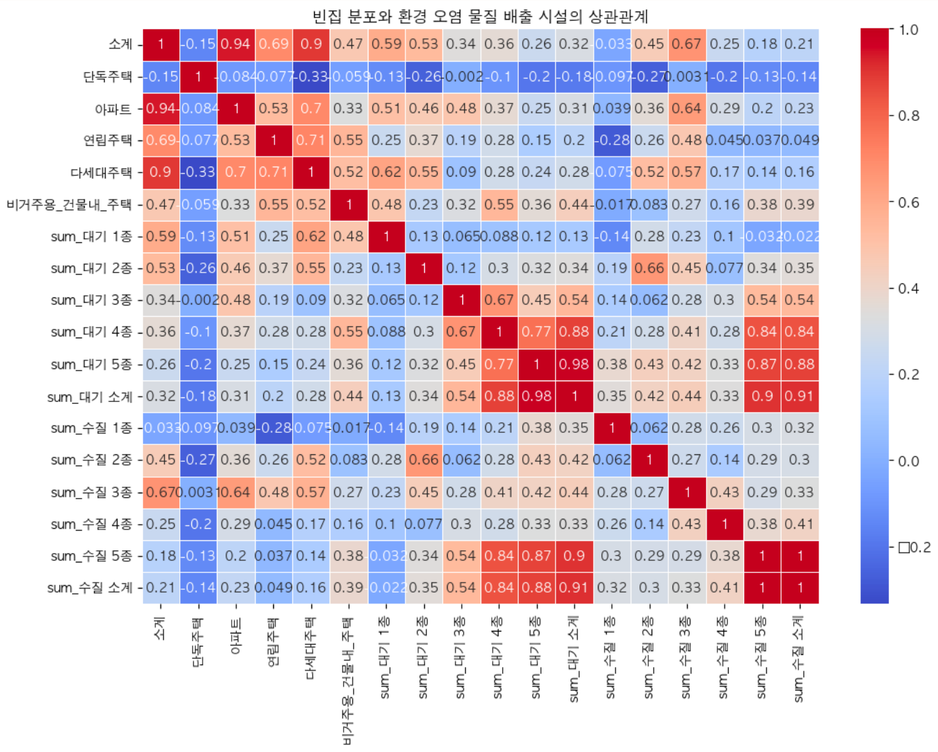
‘빈집 소계’와 ‘소음 소계’ 사이의 **높은 양의 상관관계** 확인

↓

인구 밀도가 높은 지역일 수록 소음 문제가 많이 발생

다세대 주택과 소음_소계의 상관관계 : 인구 밀도가 높은 지역일 수록 소음 문제가 심각하게 발생해 주택 공실률에 영향을 높게 미침

서울시 주택공실률과 환경오염물질 배출시설 데이터의 상관관계 분석



상관관계가 가장 높은 변수 조합

주택변수	대기오염 변수	|상관계수|	관계 종류
소계	수질 3종	0.6714	(+)
아파트	수질 3종	0.6441	(+)
다세대주택	대기 1종	0.6172	(+)
소계	대기 1종	0.5929	(+)
다세대주택	수질 3종	0.5709	(+)

‘빈집 소계’와 ‘수질 3종’. ‘대기 1종’ 사이의 **높은 양의 상관관계** 확인

↓

공실률이 높은 지역이 주로 환경 오염이 심각한 지역일 가능성이 크다는 것을 시사

산업 단지나 오염원과 가까운 곳에 위치할 가능성이 높음

상관관계 분석 결과

- 상관행렬 분석 결과, **공실률이 높은 지역**일수록 **수질 오염 및 소음 문제**가 두드러지지만
  대기오염 문제는 적다는 것을 확인

- **수질 오염과 소음 및 진동 변수**는
  “환경오염(대기/소음/수질)이 적은 행정구의 주택 공실률이 낮을 것이다” 라는 **기존 가설을 지지**

- “대기 오염이 적은 행정구의 주택 공실률이 낮다”는 가설은 분석 결과에 의해 지지되지 않음
  → 대기 오염이 심한 지역이 오히려 도심지에 위치하여 공실률이 낮기 때문

2. K-means 군집화를 통한 유사한 환경오염 패턴을 가지고 있는 행정구 군집 분석

목적 > 데이터 전처리 > K-means 클러스터링 > 실험 결과 분석

목적

- K-means 군집화를 통해 **행정구를 그룹화**하고, **각 그룹의 환경 오염 특성을 파악**하여 **유사한 환경오염 패턴을 가지고 있는 행정구 분석**
- 서울시 행정구의 유사한 환경 오염 패턴을 분석하여, **환경 오염과 주택 공실률 간의 관계 이해**

2. K-means 군집화를 통한 유사한 환경오염 패턴을 가지고 있는 행정구 군집 분석

데이터 전처리

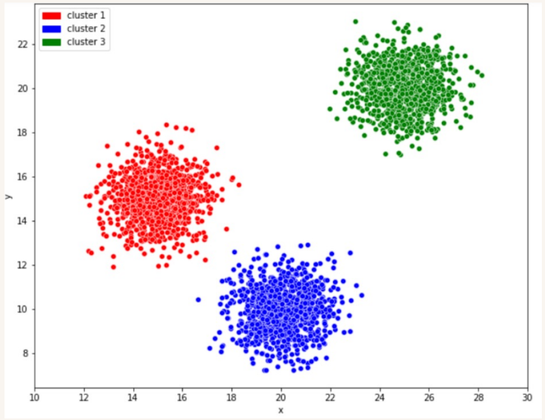
- **데이터 스케일링**
  - StandardScaler 사용
- **PCA(주성분 분석)를 사용해 2차원으로 feature 차원 축소**
  - 사용하고 있는 데이터셋의 feature 수가 총 17개이므로 '차원의 저주' 문제가 존재
  - 다중공선성 해소, 노이즈 제거, 과적합 방지 등을 위해 필요

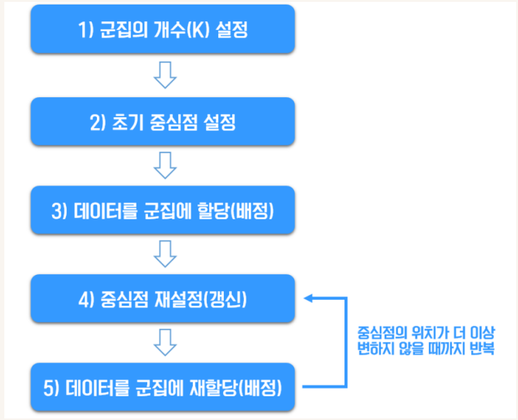
2. K-means 군집화를 통한 유사한 환경오염 패턴을 가지고 있는 행정구 군집 분석

목적 > 데이터 전처리 > **K-means 클러스터링** > 실험 결과 분석

K-means 클러스터링

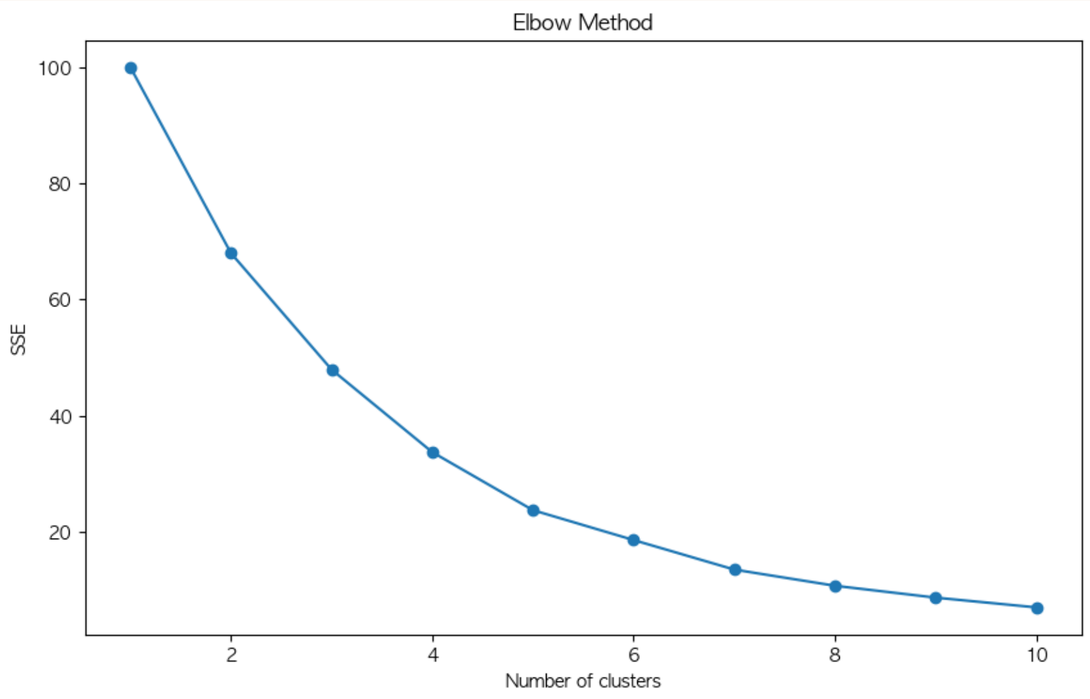
- K-means 클러스터링
  - 데이터를 유사한 특성을 가진 K개의 그룹으로 나누는 알고리즘

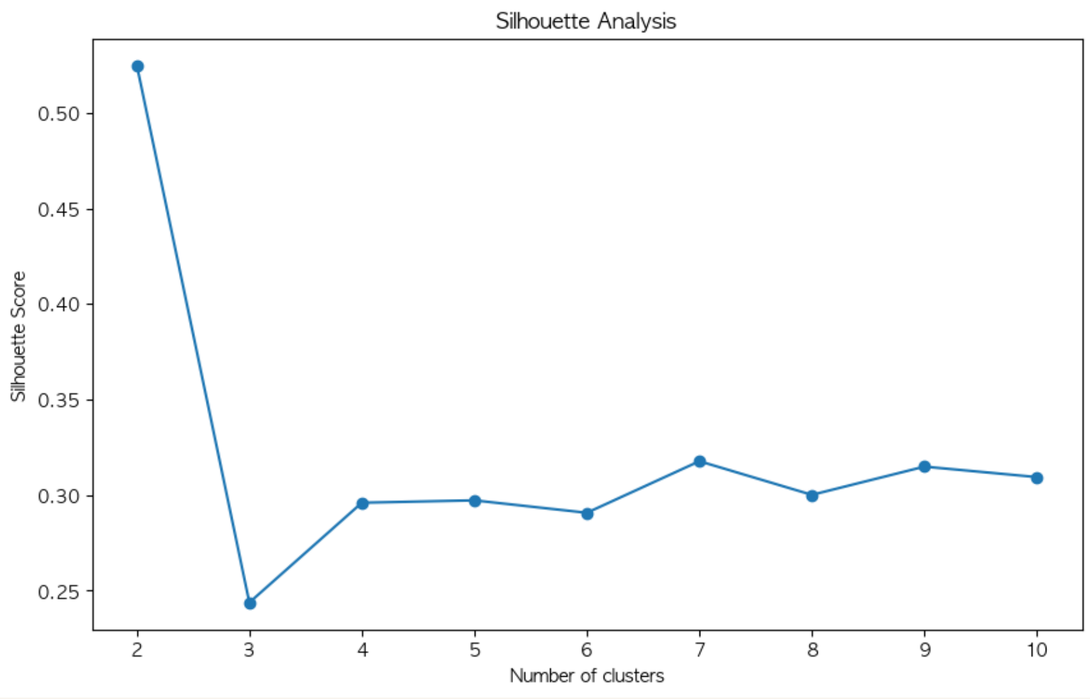




- **최적의 클러스터 수 찾기**
  - 엘보우 기법 및 실루엣 계수 분석을 사용하여 클러스터 수를 4개로 설정

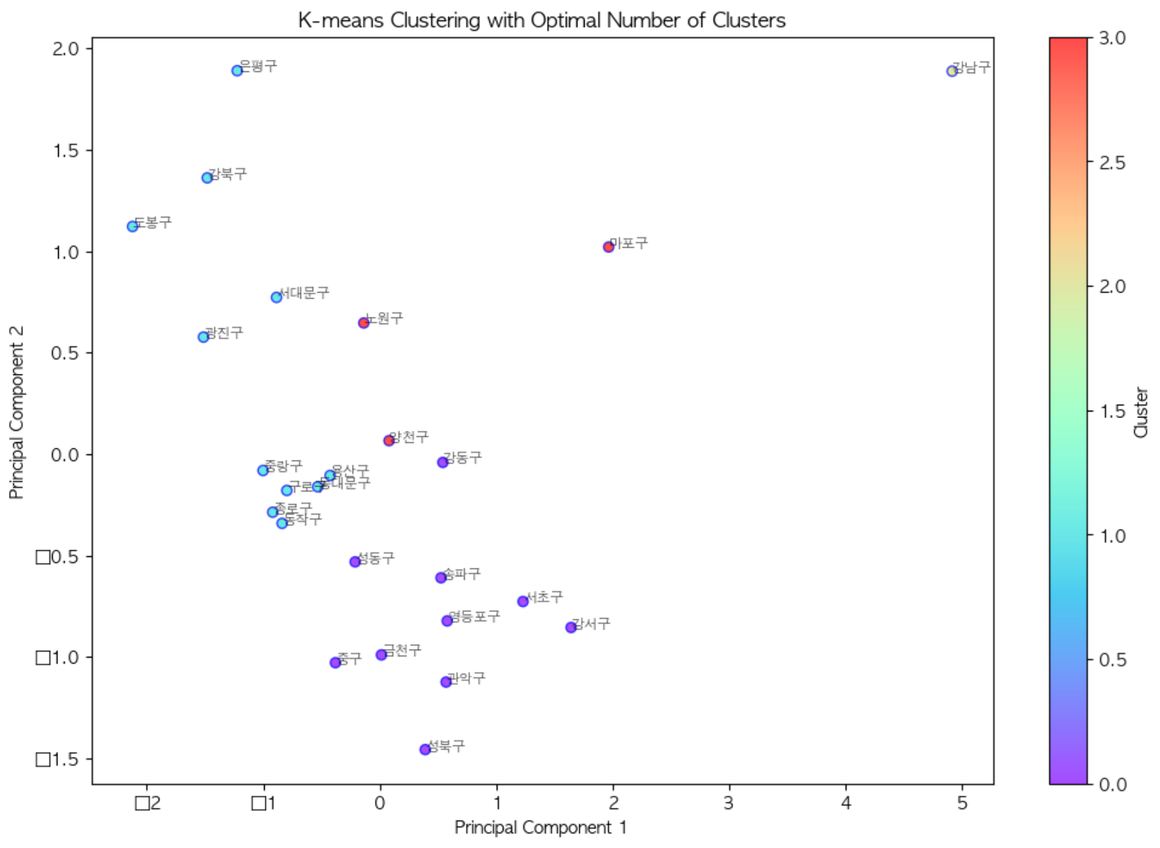
최적의 클러스터 개수 선택 - Elbow Method, Silhouette 계수





실루엣 계수와 엘보우 메서드 그래프 분석 결과,
엘보우 그래프의 기울기가 가장 급격히 줄어드는 시점인 4로 클러스터 개수 선정

K-means 클러스터링 결과

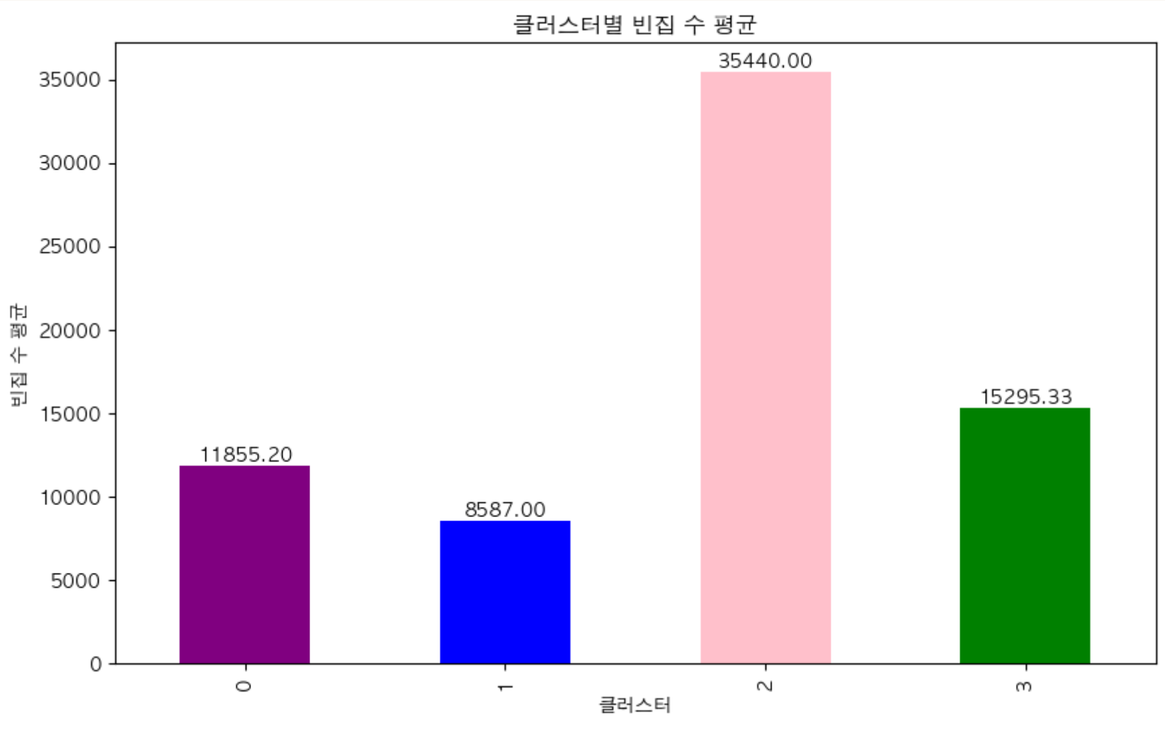


환경오염 요인과 클러스터 관계

	클러스터 0	클러스터 1	클러스터 2	클러스터3
포함 행정구	성동구, 강동구,송파구, 영등포구, 중구, 금천구, 관악구, 성북구, 서초구, 강서구	도봉구, 강북구, 은평구, 광진구, 중랑구, 용산구, 동대문구,종로구, 동작구	노원구, 양천구, 마포구	강남구
이산화질소 농도	높음 (27.00 ppm)	낮음 (22.33 ppm)	높음 (26.34 ppm)	높음 (26.11 ppm)
생활 소음	중간 (8274.60)	낮음 (6678.45)	매우 높음 (27772.00)	낮음 (5514.67)
대기 1종	낮음 (0.70)	매우 낮음 (0.27)	높음 (10.00)	중간 (8.00)
수질 3종	낮음 (8.00)	매우 낮음 (1.73)	높음 (16.00)	중간 (3.67)

- 가장 환경오염 지수가 **낮게** 측정 된 클러스터 : **클러스터 1**
- 가장 환경오염 지수가 **높게** 측정 된 클러스터 : **클러스터 2**

K-means 클러스터링 결과



클러스터 특성과 빈집 수의 관계

- **클러스터 1 과 클러스터 2의 평균 빈집 수 차이 : 75.77%**

  → 환경적 요소가 좋은 클러스터 1의 빈집 수에 비해
  환경적 요소가 좋지 않은 클러스터 2의 빈집 수가
  약 76% **더 많음**

2. K-means 군집화를 통한 유사한 환경오염 패턴을 가지고 있는 행정구 군집 분석

실험 결과 요약

- **환경 오염 요인이 높은 행정구일수록 주택 공실률이 높음**
- 높은 이산화질소 농도, 수질 오염 물질 배출량, 소음 수준이 높은 지역에서 공실률이 증가
- 이러한 실험결과는 **환경 오염이 주거 환경의 쾌적성을 저해하여 공실률 증가에 기여**한다는 가설 지지

3. 다중 회귀분석

서울시 주택 공실률 데이터와 환경오염 지수(대기·소음·수질)의
상관관계 분석을 위한 다변수 회귀분석 진행

데이터 전처리 1. 다변수 회귀분석을 위한 종속변수, 독립변수 설정

종속변수(y, label)

- **컬럼명** : 빈집_소계
- **데이터 소개**
  - 서울시 빈집 데이터셋의 지역구 별 주택 공실률 합계

종속변수 통계량 분석	
전체 합계	27400
평균(mean)	3924.41
표준편차(std)	2599.99
최솟값(min)	5371
최댓값(max)	12417

독립변수(X)

- **컬럼명** : 아황산가스농도(ppm), 이산화질소농도(ppm), 일산화탄소농도(ppm), 소음_소계, 대기_소계, 수질_소계
- **데이터 소개**
  - **서울시 지역구별 대기오염 물질 현황**
    - 아황산가스농도(ppm), 이산화질소농도(ppm), 일산화탄소농도(ppm)
  - **소음 민원 발생 현황** : 소음_소계
  - **환경오염물질 배출 시설 현황** : 대기오염물질 배출시설, 수질오염물질 배출시설

아황산가스농도(ppm)	이산화질소농도(ppm)	일산화탄소농도(ppm)	소음_소계	대기_소계	수질_소계
0.003	0.041	0.7	2016	111	206
0.004	0.045	0.9	2016	111	206
0.004	0.038	0.9	2016	111	206
0.005	0.042	1.0	2016	111	206
0.005	0.050	1.0	2016	111	206
...	...	...	...	...	...
0.003	0.040	0.7	3750	46	94
0.003	0.026	0.5	3750	46	94
0.003	0.031	0.5	3750	46	94
0.003	0.030	0.5	3750	46	94
0.003	0.039	0.6	3750	46	94

데이터 전처리 2. 데이터 스케일링

- Standard Scaler
  - Standard Scaler를 사용해서 각 컬럼의 데이터 별로 평균 0, 분산 1 의 표준 정규분포를 따르도록 독립변수 값을 정규화

$$x_{i\_}\ new = \frac{x_i - mean(x)}{stdev(x)}$$

정규화 식

데이터 전처리 3. 평가지표 선정

- RMSE, MAE, $\mathbf{R^2}$ 결정 계수 사용
  - MSE(Mean-Squared-Error)
    - 모델의 예측값이 실제값과 얼마나 일치하는 지를 나타내는 지표
    - 회귀 모델의 예측값과 실제값 간의 오차를 제곱하여 평균한 값
  - RMSE(Root Mean-Squared-Error)
    - MSE 값에 루트를 취한 값, 종속 변수와 동일한 단위를 가지기 때문에 예측 오류를 더 직관적으로 해석 가능
  - MAE(Mean-Absolute-Error)
    - 회귀모델의 예측값과 실제값 간의 오차의 절대값을 평균한 값
  - $\mathbf{R^2}$ 결정계수
    - 회귀모델의 성능을 평가하는 지표

1. 서울시 주택 공실률 - 환경 오염 지수 간 상관관계 분석

- 종속변수와 독립변수 사이에 선형관계가 존재한다고 가정
- Scikit-Learn의 LinearRegression 선형회귀 모델을 사용하여 상관관계 분석

결과 분석

MAE	1870.206
RMSE	2166.322
R^2	0.288

- R^2 값이 0.28로 **다소 낮게** 나옴
- RMSE 값 또한 **다소 높게** 나옴

→ 모델이 데이터의 분산을 잘 설명하지 못하고, 예측능력이 좋지 못함

→ 상관관계 파악에 방해가 되는 **다중공선성이 존재하는지 파악해야 함을 확인**

2. 독립변수 간 다중공선선 파악

- 독립변수 간의 다중공선성 파악을 위해 분산팽창계수(VIF)를 계산

결과 분석

feature	VIF
아황산가스농도(ppm)	14.2723
이산화질소농도(ppm)	11.5327
일산화탄소농도(ppm)	18.9365
소음_소계	4.7336
대기_소계	19.1550
수질_소계	21.2070

- **대기오염 변수** '아황산가스농도(ppm)', '이산화질소농도(ppm)', '일산화탄소농도(ppm)' 사이의 VIF가 모두 높게 나와 **다중공선성이 매우 높게 측정됨**
- **환경오염 배출시설 변수**인 '대기_소계', '수질_소계' **모두 다중공선성이 매우 높게 나옴**

→ 아황산가스농도(ppm),이산화질소농도(ppm),일산화탄소농도(ppm) VIF가 높으므로 **VIF가 낮은 이산화질소농도를 대기오염물질 관련 대표 독립변수로 사용**

2. 독립변수 간 다중공선선 파악

- 독립변수 간의 다중공선성 파악을 위해 분산팽창계수(VIF)를 계산

결과 분석

feature	VIF
아황산가스농도(ppm)	3.075
소음_소계	2.894
수질_소계	3.007

- 1차 VIF 계수 측정 결과 이후 **다중공선성이 높아보이는 변수들을 제거한 후** 재측정한 결과 **모든 독립변수의 VIF 계수가 5 미만으로 감소**
- 독립변수 간의 **다중공선성이 대부분 해소됨**을 확인

3. PCA 적용한 다중선형회귀 모델로 상관관계 분석

결과 분석

MAE	2022.854
RMSE	2406.99
R^2	0.1211

- MAE : 종속 변수의 평균 값과 비교했을 때(3924.41) 비교적 큰 오차 결과를 보임
- RMSE : 모델의 예측 오차가 비교적 큰 편임
- R^2 : 모델이 데이터 변동의 약 12.11%만을 설명하고, 모델이 종속 변수의 변동성을 거의 설명하지 못하고 있음

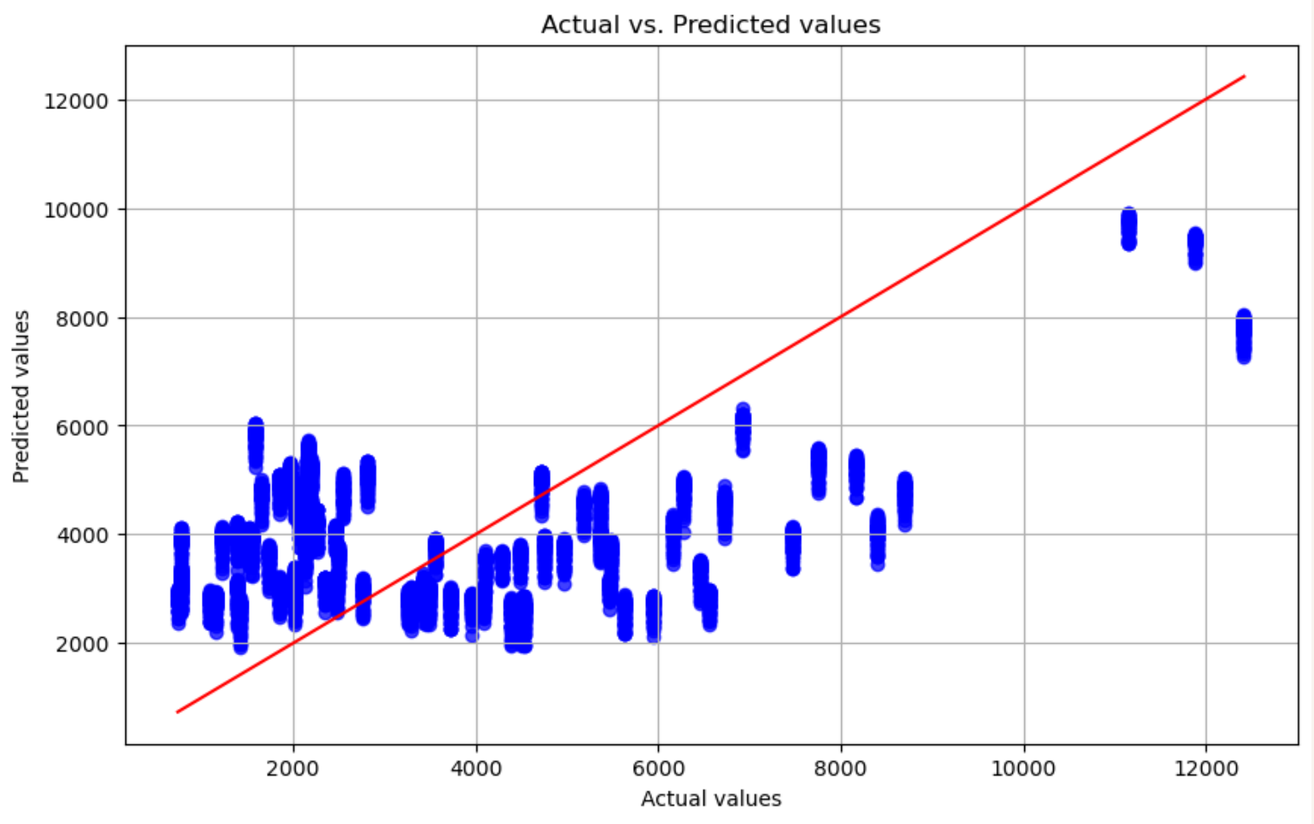
→ MAE와 RMSE 값이 크고, R^2 값이 낮은 것으로 보아 **모델의 예측 성능은 낮은 편**

빅데이터 분석 기말 프로젝트

3. PCA 적용한 다중선형회귀 모델로 상관관계 분석

실제값과 모델 예측값 시각화 그래프 분석

- 대부분의 데이터 포인트가 회귀모델(빨간색 대각선)에 크게 벗어나 있음
- 모델이 실제 값을 잘 예측하지 못함



결론

선형 회귀 모델으로는 현재 데이터의 비선형적 패턴을 잘 포착할 수 없기 때문에
다변수 비선형 회귀분석 방법인 랜덤포레스트를 사용하기로 함

4. 랜덤포레스트를 사용해 비선형 관계 파악

- 다변수 회귀분석에서의 차원의 저주 문제를 해결하기 위해 주성분 분석으로 차원 축소 진행
- 랜덤포레스트 회귀분석으로 주택 공실률과 환경오염지수 간의 비선형 관계 모델링 진행

결과 분석

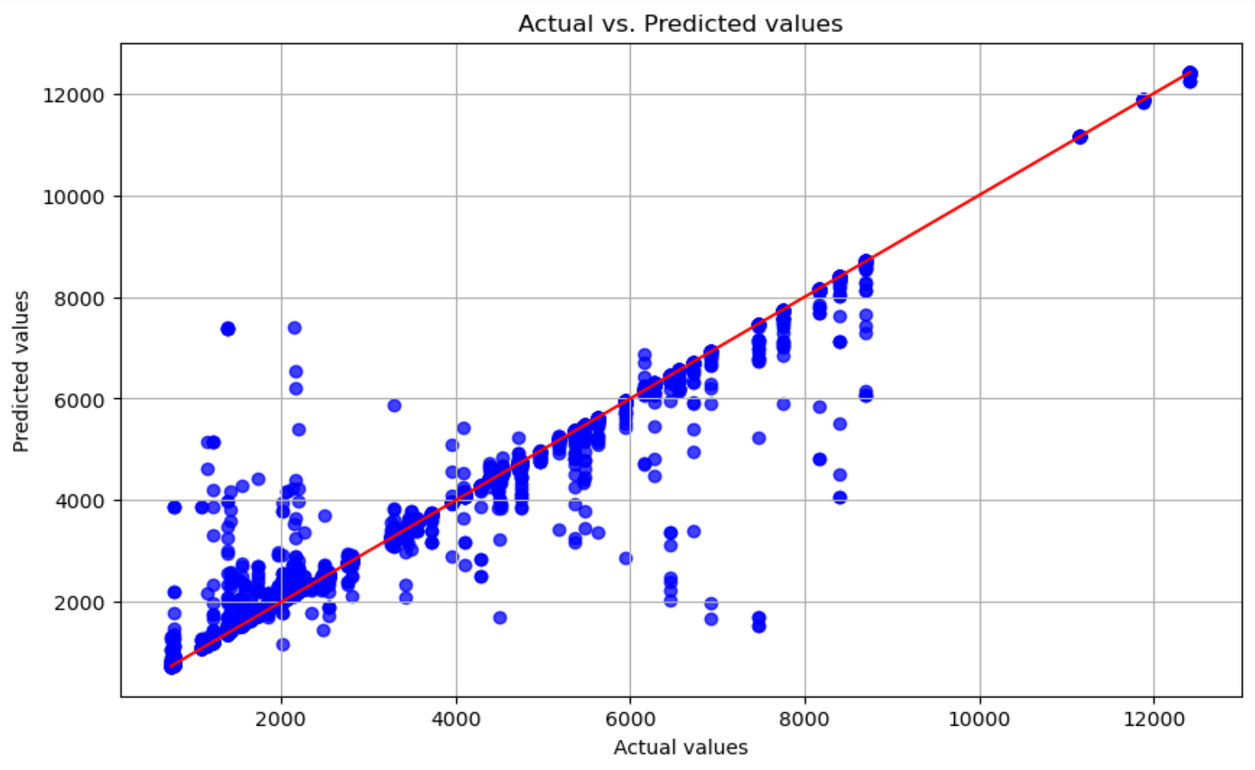
MAE	94.9163
RMSE	452.59
Adjusted R^2	0.96

- **MAE**
  - 종속 변수의 평균값과 비교했을 때, MAE는 평균 값의 약 2.42%에 해당
  - 모델의 예측이 실제값에 매우 가까움을 의미
- **RMSE**
  - 종속 변수의 평균값이 3,924.41이고 표준편차가 2,599.99인 점을 고려할 때, RMSE는 평균 값의 약 11.53% 에 해당
  - 모델의 예측 오차가 데이터의 전체 변동성에 비해 작음을 의미
- **Adjusted R^2 Score**
  - 모델이 종속 변수의 변동성을 약 96.89% 설명하고 있음을 확인할 수 있고 모델이 데이터의 분산을 매우 잘 포착하고 있음을 의미

4. 랜덤포레스트를 사용해 비선형 관계 파악

Actual vs. Predicted Plot 산점도 그래프

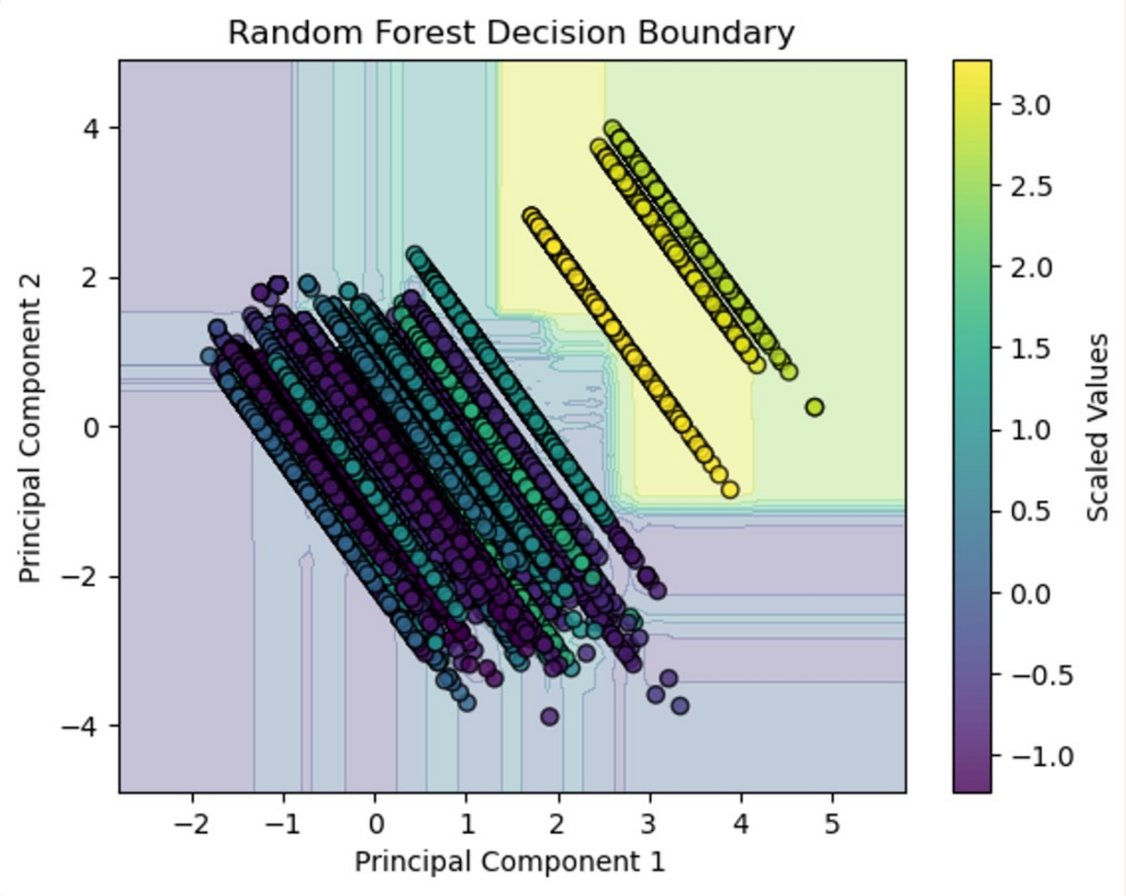
- 실제 값과 예측 값을 비교한 산점도 그래프에서 대부분의 데이터 포인트가 회귀모델(빨간색 대각선)에 가까이 위치
- 모델이 실제 값을 매우 잘 예측하고 있음을 확인



4. 랜덤포레스트를 사용해 비선형 관계 파악

결정 경계 시각화 그래프 분석

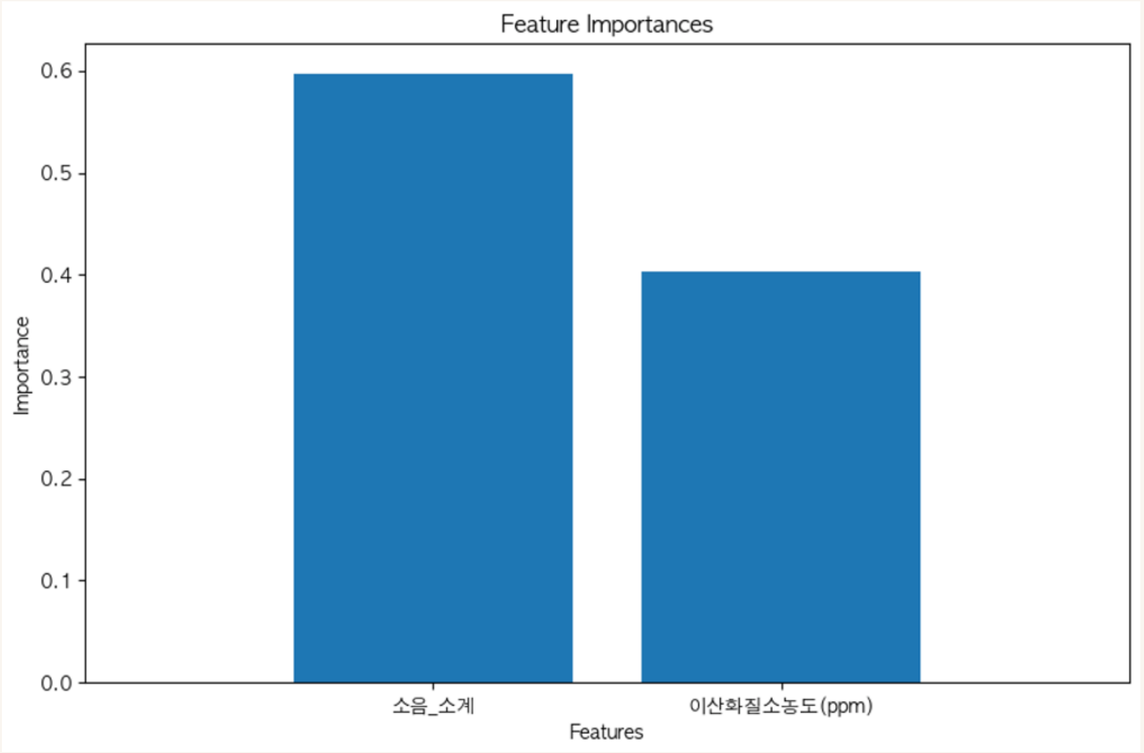
- 그래프에서 각 데이터 샘플이 여러 색상영역에 걸쳐 분포되어 있음
  → **현재 데이터는 복잡한 비선형적 결정경계를 형성**
- 대부분의 데이터 샘플들이 같은 색상영역에 몰려 있는 것을 확인 가능
  → 모델이 해당영역의 데이터에 대해서 **일정한 예측값을 생성하여 데이터를 효과적으로 분할하고 있음**



5. 랜덤포레스트 모델을 사용한 특성 중요도 추출 및 시각화

결과 분석

- 행정구 별 주택 공실률에 가장 큰 영향을 미치는 환경오염관련 변수는 소음_소계 와 이산화질소농도 임을 확인



5. 결론 및 프로젝트 의의

결론

1. 가설 ‘환경오염(대기·소음·수질)이 적은 행정구의 주택 공실률이 낮을 것이다.’ 검증 결과
   - 소음 및 수질 오염이 심할 수록 주택 공실률이 높다는 것을 확인 (가설 지지)
   - 대기 오염물질 중, 이산화질소 농도가 주택 공실률에 가장 큰 영향을 미침 (가설 지지)

2. 유사한 환경오염 패턴을 가지고 있는 행정구 군집 분석 결과
   - 환경오염 지수가 높은 ‘클러스터 2’에 속해 있는 행정구의 빈집 수가 더 많다는 것을 확인
   - 높은 이산화질소 농도와 생활 소음이 실제 빈집 증가의 원인으로 작용함을 확인

프로젝트 의의

환경오염이 주거지 선택에 미치는 영향을 식별하고, 서울시 환경오염 문제 및 주거 만족도 개선에 기여하고자 함

- 환경오염에 따른 서울시 주택공실률 증가 요인을 줄이기 위해 소음 차단 시설 설치, 수질 개선을 위한 오염물질 배출 규제 등의 정책적 노력이 필요함을 제시
- 추후 경제적 요인, 교통 접근성 등 주택공실률에 영향을 미칠 수 있는 다른 변수들을 고려하여 더 포괄적인 분석 가능

Thank You