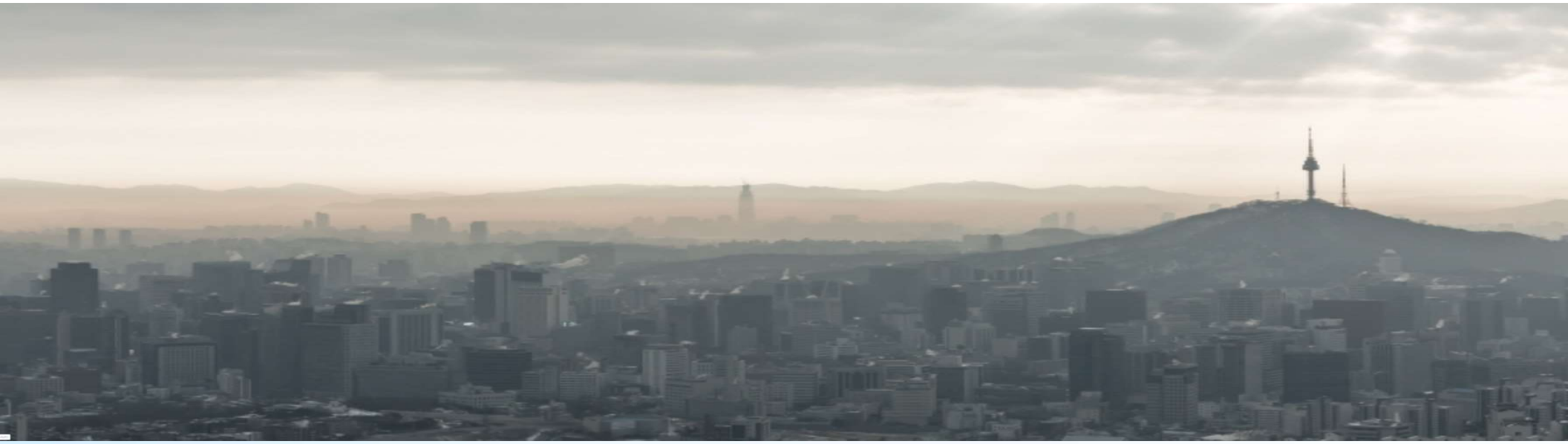


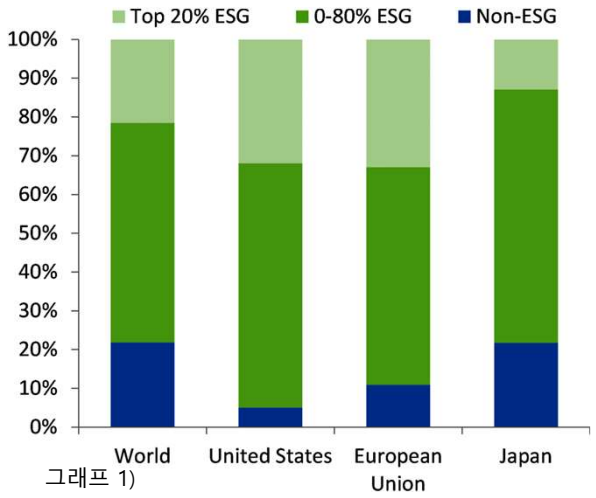
미세먼지 영향인자 분석 및 개선안 도출

C반 김민정

추진배경 및 현황

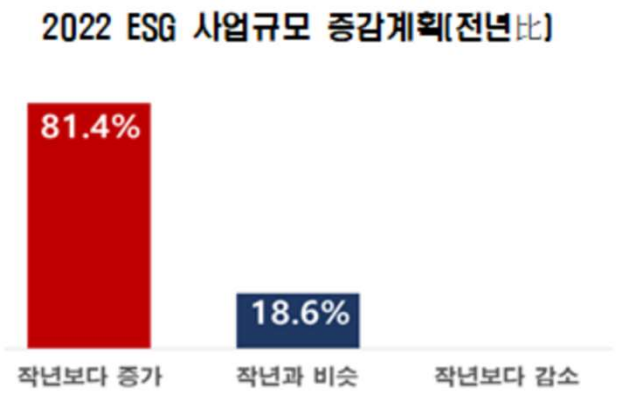
- **분석 배경**
- 세계적으로 환경, 기후변화 문제가 중요해짐에 따라 ESG 경영과 같은 기업의 사회적 책무가 거론됨
- 이에 따라 매년 사회적 이슈로 떠오르는 미세먼지 문제에 대해 각 산업별 해결책을 강구하고자 함

- **분석 방향**
- 미세먼지 주요 영향인자 분석
- 해당 주요 영향인자와 기업 간의 상관관계
- 해당 주요 영향인자와 각 산업 군별 상관관계



그래프 1)

세계적으로 80% 기업이 ESG 경영에 동참



그래프 2)

매출액 300대 기업 81%가 ESG 사업예산 더 늘릴 전망

미세먼지 정의

- **미세먼지란?**

- 대기 중에 떠다니거나 흩날려 내려오는 입자상물질로, 입자의 지름이 10㎛ 이하인 먼지(PM-10, 미세먼지)와 지름이 2.5㎛ 이하인 먼지(PM-2.5, 초미세먼지)로 구분

- **미세먼지 발생원**

- 자연적 발생원 : 흙먼지, 바닷물에서 생기는 소금, 식물의 꽃가루 등

- 인위적 발생원 : 화석연료를 사용할 때 생기는 매연, 자동차 배기가스, 건설현장에서 발생하는 날림먼지, 공장 내 분말형태의 원자재 · 부자재, 취급공정에서의 가루성분, 소각장 연기 등

- **미세먼지 1차/2차 발생원**

- 굴뚝 등 여러 발생원에서 고체 상태의 먼지로 나오는 경우(1차 발생원)와 이러한 1차 발생원에서 가스 상태로 나온 물질이 공기 중에서 다른 물질과 화학반응을 일으켜 미세먼지가 되는 경우(2차 발생원)로 나누어짐

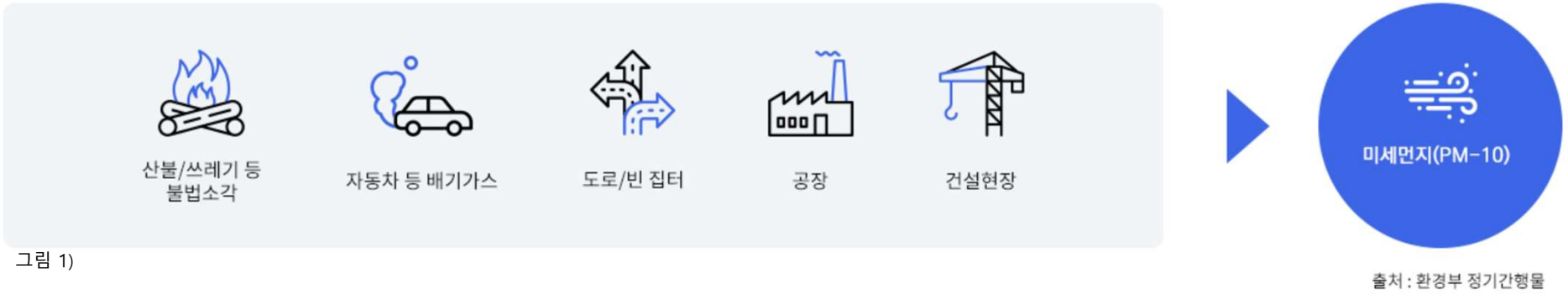


그림 1)

출처 : 환경부 정기간행물

분석 계획

목적	분석방법	주요내용	담당자	일정	비고
데이터 특성 파악	기초통계 분석	데이터의 타입, 평균, 표준편차, 최소값, 최대값 등의 기본 정보 확인	김민정	8/14	
	결측치 분석	데이터에 결측치가 있는지 확인 후 처리	김민정	8/14	
전체 데이터의 분포 특성 및 변수 간의 관련성 확인	Histogram 분석	전체 데이터의 구간별 분포 현황과 패턴 확인	김민정	8/14	
	Box Plot 분석	Box plot을 이용하여 데이터에 이상치가 있는지 확인 후 처리	김민정	8/14	
	산점도 분석	목표변수와 설명변수 간 분포, 패턴, 관계 확인	김민정	8/14	
	상관 분석	목표변수와 설명변수 간 양/음의 상관관계 파악	김민정	8/14	
	Heatmap 분석	상관 분석 진행한 데이터로 시각화하여 조합별 비교	김민정	8/14	
잠재인자 (Vital Few) 탐색	회귀 분석	회귀분석을 통하여 잠재인자 도출, 다중공선성 확인	김민정	8/14	
	Decision Tree	Decision Tree를 통해 변수 별 중요도 확인 후 잠재인자 도출	김민정	8/14	
	Random Forest	Random Forest를 통해 변수 별 중요도 확인 후 잠재인자 도출	김민정	8/14	
	Gradient Boosting	Gradient Boosting을 통해 변수 별 중요도 확인 후 잠재인자 도출	김민정	8/14	
예측 모델링 진행	회귀 분석	도출된 잠재인자를 통하여 예측 회귀모델 생성	김민정	8/14	
	Decision Tree	도출된 잠재인자를 통하여 예측 DT모델 생성	김민정	8/14	
	Random Forest	도출된 잠재인자를 통하여 예측 RF모델 생성	김민정	8/14	
	Gradient Boosting	도출된 잠재인자를 통하여 예측 GB모델 생성	김민정	8/14	
모델 평가	오차함수/Bar chart	Bar chart를 통해 mse,rmse,mae,mape 확인 및 최종 모델평가	김민정	8/14	
추가 데이터 분석	기초통계/Pie chart	위 과정을 통해 도출된 Vital few와 산업군 간 영향력 분석	김민정	8/14	

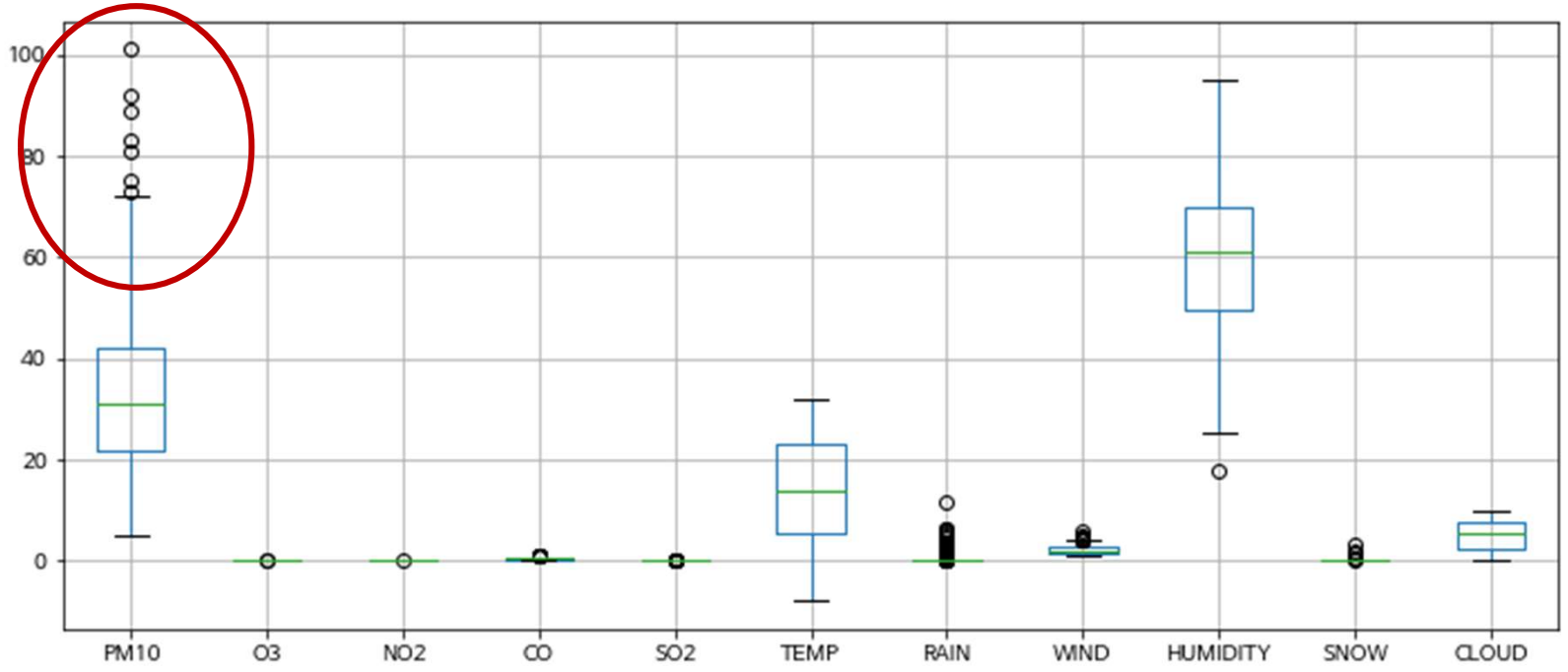
데이터 현황

- **결측치 처리**

- PM10: 목표변수로 대체가 어려워 제외 처리
- CO: 산점도 확인 결과 목표변수(PM10)와 주요한 관계에 있을 것으로 예상되어 중앙값으로 대체

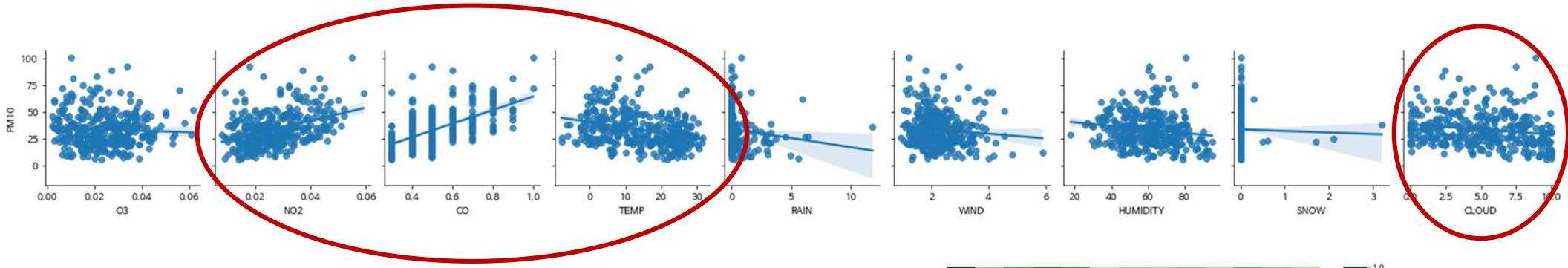
- **이상치 처리**

- PM10: IQR 최대값을 넘긴 이상치가 보이나, 특정 환경에서 관측될 수 있는 부분으로 생각하여 이상치 처리하지 않음
- (ex) 경험상 봄~겨울에 고농도 미세먼지가 많이 발생하는 경향을 보이며, 이상치를 제외하지 않았을 때 해당 관계 역시 분석이 가능할 것으로 예측

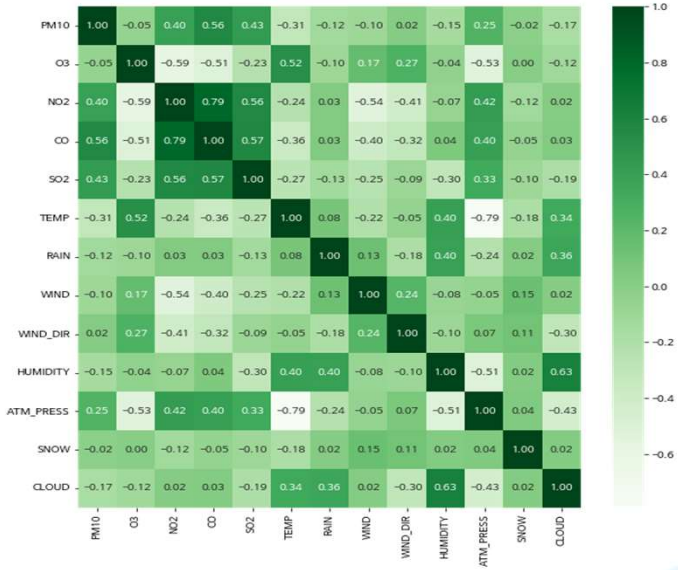


탐색적 분석

- **Scatter Matrix를 통한 잠재인자 확인**
- NO2: (+) 양의 비례 관계 확인
- CO: 강한 (+) 양의 비례 관계 확인
- TEMP: (-) 음의 관계 확인
- CLOUD: 약한 (-) 음의 관계 확인



- **상관분석 / Heatmap을 통한 잠재인자 확인**
- CO > SO2 > NO 순으로 (+) 양의 상관관계가 있으며
- TEMP와는 (-) 음의 상관관계가 있음을 확인함



Vital Few 도출

- **선택 변수**

- CO/O3/TEMP : 4개의 모델에서 모두 공통으로 선택된 변수
- CLOUD/WIND_DIR: 3개의 모델에서 공통으로 선택된 변수
- NO2/WIND: 2개의 모델에서 공통으로 선택된 변수

모델 종류	다중선형회귀분석	Decision Tree	Random Forest	Gradient Boosting
예측률	44.4%	100%	93%	88.9%
중요변수	O3	CO	CO	CO
	NO2	O3	O3	O3
	CO	TEMP	TEMP	TEMP
	TEMP	CLOUD	CLOUD	WIND_DIR
	WIND_DIR	WIND	WIND_DIR	CLOUD
	ATM_PRESS	ATM_PRESS	WIND	NO2

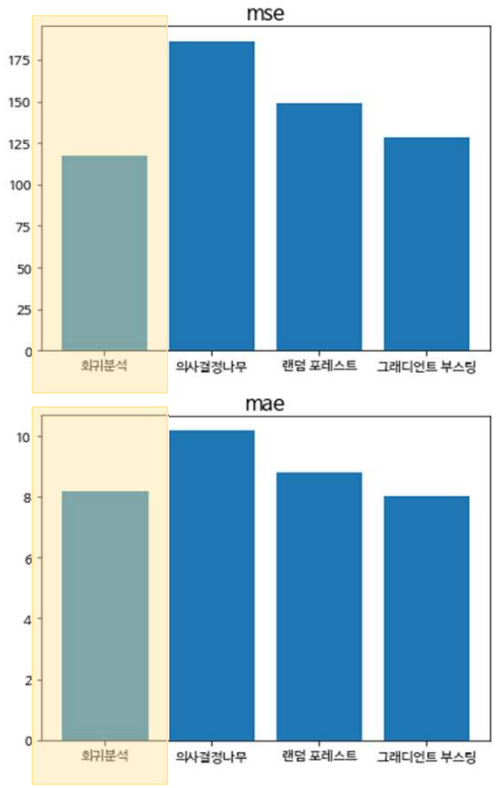
*회귀분석 진행 시 설명변수 간 다중 공선성은 나타나지 않아 따로 고려하지 않음

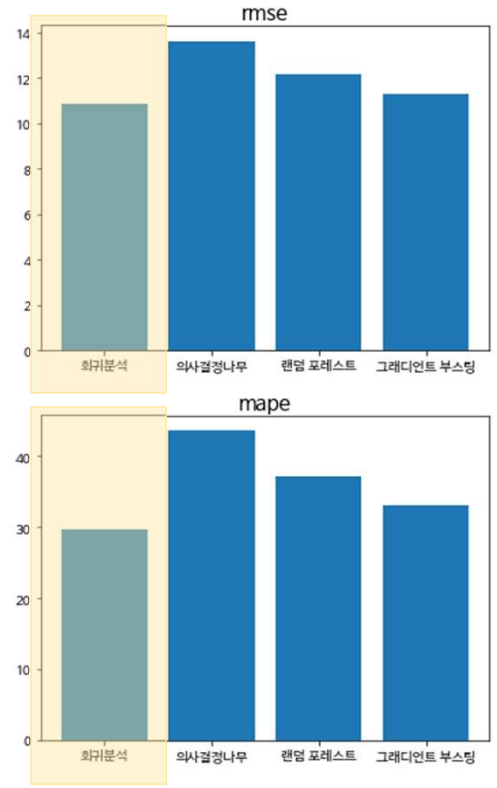
모델링 & 요약

- **모델간 예측률 비교**

- 다중선형회귀분석 모델이 48.8%로 가장 높은 정확도를 보여 최종 모델로 선정
- R-square에 따라 모델의 설명도는 49.5%로 다소 낮지만, F-statistic 통계량은 유의수준 (0.05)보다 낮아 모델 자체는 유의함
- 해당 모델에 따르면 변수 CLOUD는 유의하지 않은 변수임을 알 수 있다.

모델 종류	다중선형회귀분석	Decision Tree	Random Forest	Gradient Boosting
예측률	48.8% (R)	31.8%	42.8%	48.4%





OLS Regression Results

Dep. Variable:	PM10	R-squared:	0.495
Model:	OLS	Adj. R-squared:	0.484
Method:	Least Squares	F-statistic:	43.60
Date:	Mon, 15 Aug 2022	Prob (F-statistic):	1.92e-48
Time:	06:26:07	Log-Likelihood:	-1403.3
No. Observations:	365	AIC:	2825.
Df Residuals:	356	BIC:	2860.
Df Model:	8		
Covariance Type:	nonrobust		

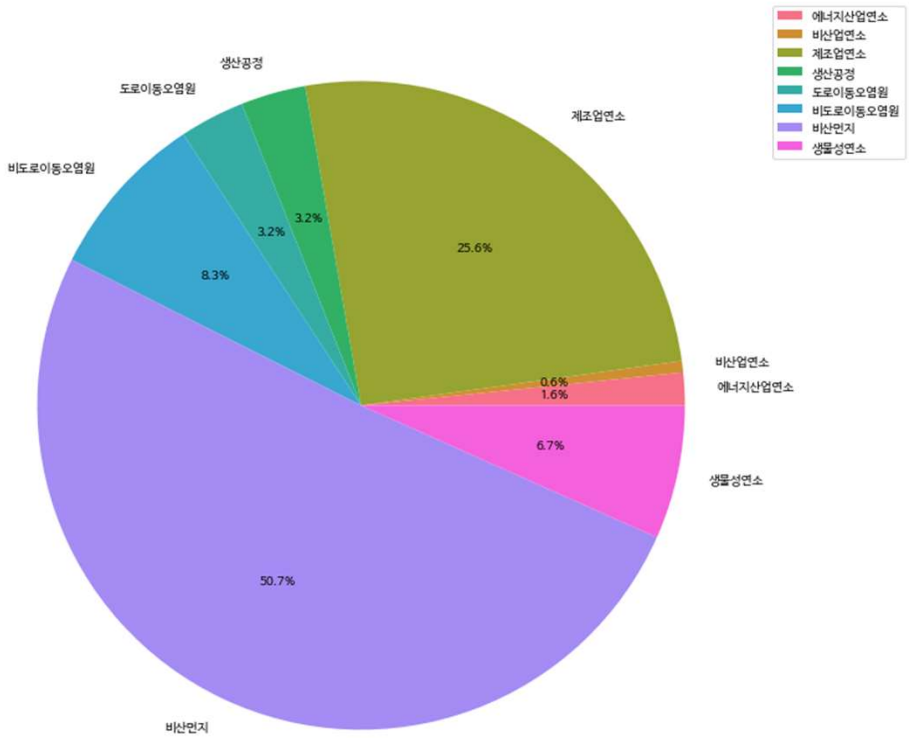
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	307.2840	150.132	2.047	0.041	12.028	602.540
O3	612.4717	78.825	7.770	0.000	457.450	767.493
CO	65.4300	7.546	8.671	0.000	50.590	80.270
NO2	508.0278	122.590	4.144	0.000	266.937	749.118
TEMP	-0.5621	0.125	-4.495	0.000	-0.808	-0.316
WIND_DIR	0.0384	0.010	3.834	0.000	0.019	0.058
ATM_PRESS	-0.3380	0.148	-2.288	0.023	-0.629	-0.047
CLOUD	-0.2412	0.261	-0.924	0.356	-0.755	0.272
WIND	2.4212	1.131	2.141	0.033	0.197	4.645

Omnibus:	96.893	Durbin-Watson:	1.196
Prob(Omnibus):	0.000	Jarque-Bera (JB):	263.658
Skew:	1.247	Prob(JB):	5.59e-58
Kurtosis:	6.335	Cond. No.	2.70e+05

모델링 & 요약

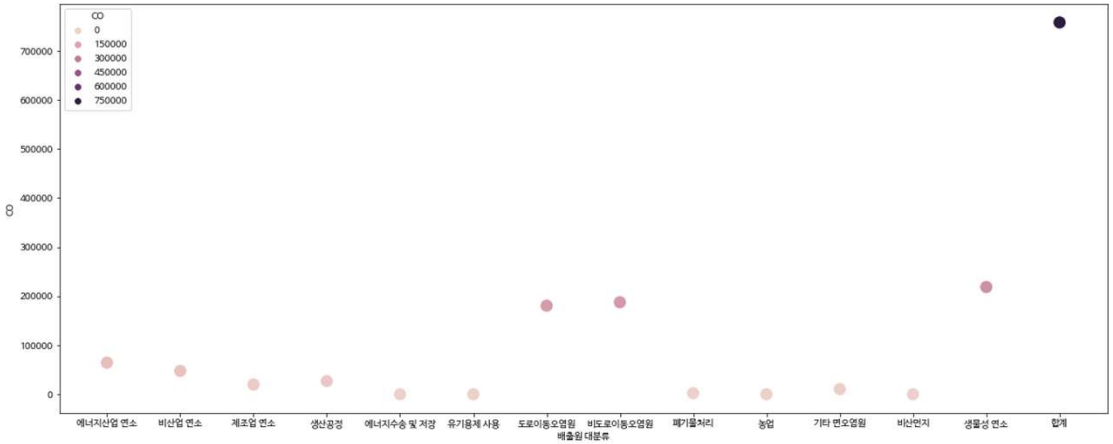
- **추가 데이터 분석**

- 앞선 분석을 통하여 Vital few로 CO/O3/TEMP/NO2/WIND/WIND_DIR/ATM_PRESS 를 도출함
- 이러한 영향인자들이 실제로 각 산업군과 어떻게 연관이 있는지 확인해보기 위하여 추가 데이터 분석 진행함
- Data는 국가미세먼지정보센터의 2019년 배출량 통계 이용함

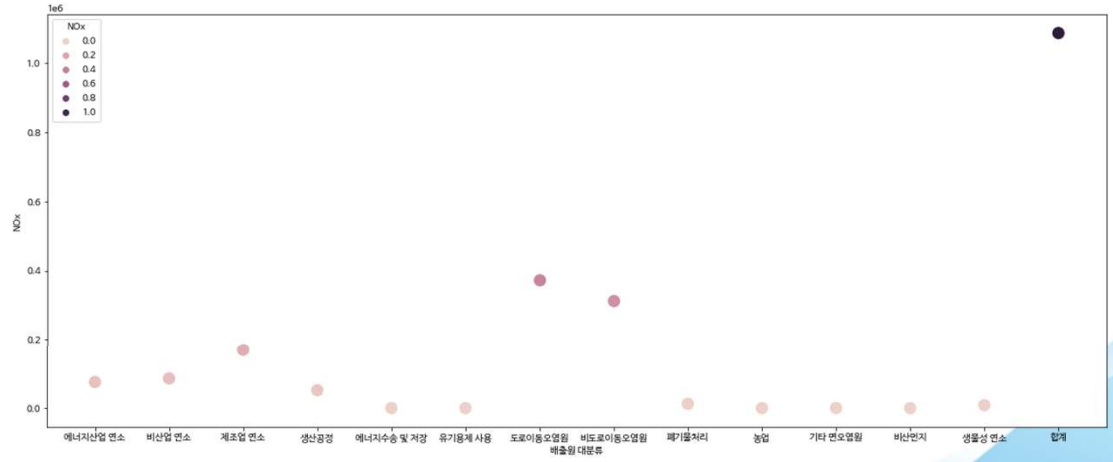


PM10 주요 배출원: 비산먼지>제조업연소>비도로이동오염원

CO 주요 배출원: 생물성 연소> 비도로이동오염원 > 도로이동오염원



NO 주요 배출원: 도로이동오염원>비도로이동오염원>제조업연소



결론/대안제시

▪ 분석결과

- PM10 농도의 경우 CO, O3, TEMP, NO2, WIND, WIND_DIR, ATM_PRESS와 상관관계가 있음을 알 수 있다.
- PM10의 경우 비산먼지 > 제조업연소 > 비도로이동오염원 순으로 주요 배출원임을 알 수 있다.
- CO의 경우 생물성 연소 > 비도로이동오염원 > 도로이동오염원 순으로 주요 배출원임을 알 수 있다.
- NO의 경우 도로이동오염원 > 비도로이동오염원 > 제조업연소 순으로 주요 배출원임을 알 수 있다.
- TEMP, WIND, WIND_DIR, ATM_PRESS의 경우 기업의 생산활동과 직접적인 연관성을 찾지 못하여 추가 분석이 필요하다.
- 다만, 해당 분석을 통해 산업군 중 건설업, 제조업, 축산업 등이 미세먼지 영향인자와 밀접 관련이 있음을 알 수 있다.

▪ 배출원 분류

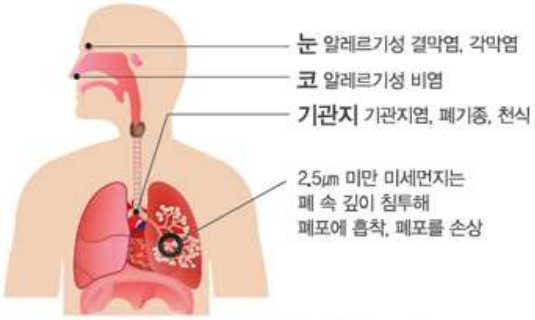
배출원 대분류	배출원 중분류
에너지산업 연소	공공발전시설, 지역난방시설, 석유정제시설, 민간발전시설
비산업 연소	상업 및 공공기관 시설, 주거용시설, 농업·축산·수산업시설
제조업 연소	**연소시설, 공정로, 기타**
생산공정	석유제품, 산업, 제철제강업, 무기화학제품제조업, 유기화학제품제조업, 목재・펄프제조업, 식음료가공, 기타제조업
에너지수송 및 저장	휘발유공급
유기용제 사용	도장시설, 세정시설, 세탁시설, 기타 유기용제사용
도로이동오염원	**RV, 승용차, 택시, 승합차, 버스, 화물차, 특수차, 이륜차**
비도로이동오염원	**철도, 선박, 항공, 농업기계, 건설장비**
폐기물처리	폐기물소각, 기타 폐기물 처리
농업	분뇨관리, 비료사용농경지
기타 면오염원	산불 및 화재, 동물
비산먼지	**건설공사, 나대지, 농업활동, 도로재비산먼지, 비포장도로 비산먼지, 축산활동, 폐기물처리, 하역 및 야적**
생물성 연소	고기 및 생선구이, 노천 소각, 농업잔재물 소각, 목재 난로 및 보일러, 숯가마, 아궁이

결론/대안제시

▪ **미세먼지 영향**

- 앞서 도출된 미세먼지 영향인자는 아래와 같이 인체에 악영향을 주어 사회적 문제를 야기함

그림 2) 〈미세먼지의 상대적 크기 비교〉



〈미세먼지가 인체에 미치는 영향〉

▪ **개선안 도출**

- 미세먼지 영향인자와 밀접 관련한 산업: 기업 내부에서 미세먼지 저감 대책 강구함으로써 ESG 경영 성과 향상 기대
- 하기 타사 사례를 참고하여 벤치마킹하는 것이 필요

*신형 집진기가 설치된 슬래그 포트에 슬래그를 따라내는 과정
그림3)

❖ **포스코: 환경설비 투자**

성과: 미세먼지 배출을 기존 比 최대 90%까지 저감 가능

- 배가스 청정브리더 신규 적용
- 배관 설비 내 먼지 포집 설비 추가
- 하강기류를 생성해 집진 효율을 높이는 집진 신기술 개발

- ✓ 기존에 해왔던 데이터 분석과 달리 주요 영향인자부터 개선안 도출까지 진행하는 것은 처음이라 막막하기도 했으나, 그래도 step by step으로 배웠던 기법들을 이용해 보려 했습니다.

- ✓ 이번 분석을 통해서 실제 데이터 분석 진행 전 데이터 분석에 대한 시나리오 개괄을 짜는 것이 중요하다고 느꼈습니다.

- ✓ 또한, 모델링 함에 있어 회귀분석이나 ML 모델에 대해 깊이 있는 이해를 바탕으로 분석하여야 정확한 결과를 도출 할 수 있는 것 같습니다.

- ✓ 주어진 데이터를 바탕으로 미세먼지 농도와 주요 영향인자 간의 상관분석 뿐만 아니라, 이를 바탕으로 각 산업군에서는 미세먼지 농도에 어떠한 영향을 미치는지에 대해 통계적으로 분석해보고 싶었으나 파생변수에 대한 지식이 부족함을 느꼈습니다.

- ✓ 통계 기초지식이 부족하여 발생한 문제인 것 같아 추후 통계 공부가 더 필요함을 느꼈고, 도메인 지식 역시 중요함을 배웠습니다.

출처

- 그래프 1: https://www.oecd-ilibrary.org/sites/e9ed300b-en/index.html?itemId=/content/component/e9ed300b-en
- 그래프2: http://www.fki.or.kr/FkiAct/Promotion/Report/View.aspx?content_id=610cfd06-8a31-4874-a275-40bdc333bdd6&cPage=&search_type=0&search_keyword=
- 그림 1: https://www.air.go.kr/index.do
- 그림 2: https://www.air.go.kr/index.do
- 그림 3: https://www.posco.co.kr/
- 추가 데이터: https://www.air.go.kr/index.do