미니프로젝트

대기 오염도와 태양광 발전량의 상관관계 분석

4조

김민영

김영남

목 차

1. 프로젝트 개요

1.1 프로젝트 기획 배경 및 목표

(1) 배경 : R을 활용하여, 친환경에너지인 태양광 발전량과 서울시의 대기오염도와의 상관관계를 분석함

- 이를 통해 국가에서 진행되는 친환경 에너지 사업에 세금낭비를 방지하고자, 투자대비 고효율을 내기위해 적절한 지역을 찾아보고자 함.

(2) 목표 : 서울시 대기오염도와 서울시 태양광 발전량을 분석해 상관관계를 찾으며 서울보다 효율이 좋은 지역 결과를 도출

1.2 구성원 및 역할

이름	전공	역할	구현부분
김민영	통계학과	데이터 수집/가공 분석 및 시각화	데이터 가공 및 시각화 회귀분석
김영남	산업공학과	데이터 수집 분석 및 시각화	데이터 시각화 상관 관계 분석

2. 프로젝트 결과

2.1 데이터 수집

- 2018년 서울시 일별평균대기오염도 (csv 파일 크롤링)
- 2018년 서울시 한국전력거래소 기준 태양광발전량 (csv 파일 크롤링)
- 2018년 전국 이산화질소 발생량 (csv 파일 크롤링)
- 2018년 전국 오존 발생량 (csv 파일 크롤링)

2.2 데이터 전처리

2.2.1 결측값 탐색

<오염도 데이터>

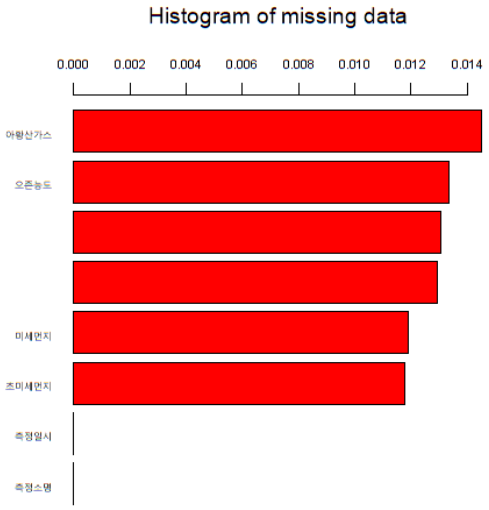
A	B	C	D	E	F	G	H
측정일시	측정소명	이산화질소농도	오존농도	이산화탄소농도	아황산가스	미세먼지	초미세먼지
20180101	강남구	0.033	0.01	0.6	0.006	34	22

오염도 데이터의 변수명과 첫번째 행이다. 미세먼지와 초미세먼지의 단위는 ㎍/㎥이고, 그 외의 측정 단위는 PPM이다.

```
FALSE   TRUE
61201    599
```

결측값의 개수는 총 61800개의 관측치 중 559개로 약 0.9%이다.

먼저 오른쪽의 그래프는 오염도 데이터를 활용한 결측값 히스토그램이다. 측정일자와 측정소명은 결측값이 존재하지 않지만 그 외의 나머지 변수들은 약 0.01%로 결측값을 포함하고 있다.



결측값이 한 변수에 치우쳐지지 않고, 변수별로 비슷하게 분포하고 있기 때문에 제거해고 분석해도 큰 문제가 되지 않지만, 오차의 최소화를 고려하여 변수별 평균값으로 결측값을 대체하고 진행하였다.

<에너지 데이터>

A	B	C	D	E
연료원	지역세부구분	거래일	시간	전력거래량
태양광	서울시	2018-01-01	5	0
태양광	서울시	2018-01-01	6	0
태양광	서울시	2018-01-01	7	0
태양광	서울시	2018-01-01	8	0.00254
태양광	서울시	2018-01-01	9	0.590215
태양광	서울시	2018-01-01	10	2.847764

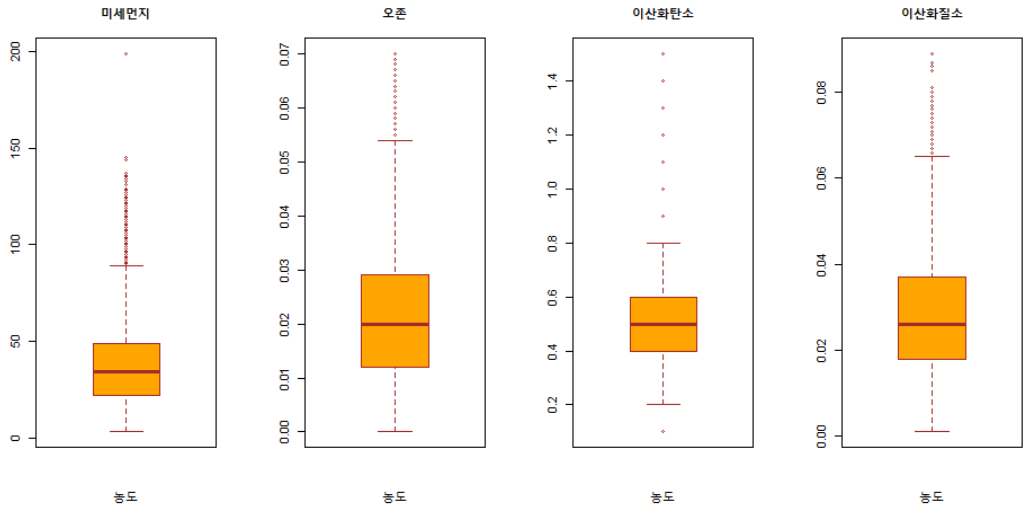
에너지 데이터의 일부이다. 지역은 서울시를 기준으로 추출했고, 거래일은 2018년 시간은 1시간 단위로 측정된 것을 확인할 수 있다. 전력이 사용되지 않은 시간대는 거래량이 0으로 측정되었다.

```
  FALSE
113160
```

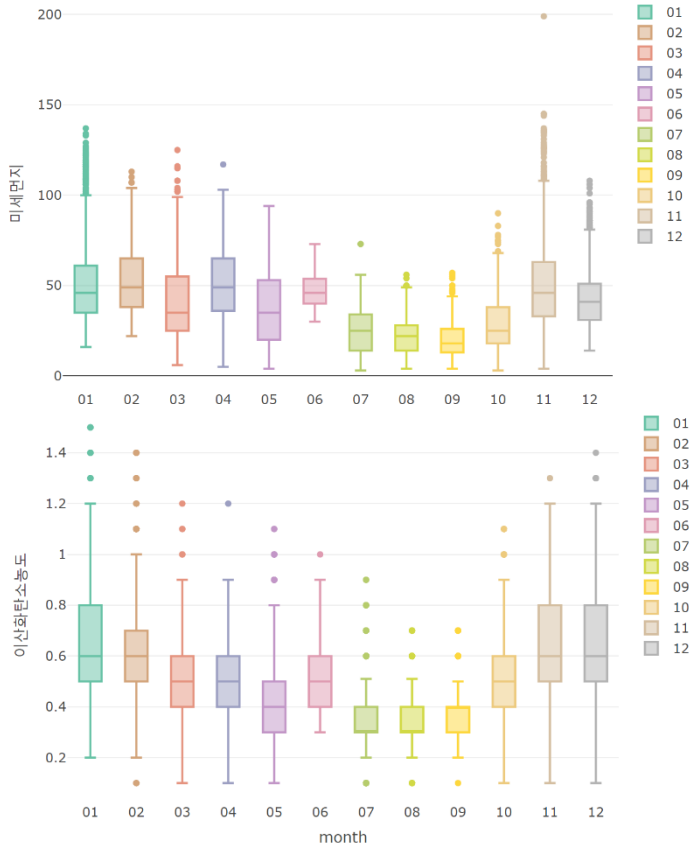
결측값은 존재하지 않는다.

2.2.2 이상치 탐색

<오염도 데이터>

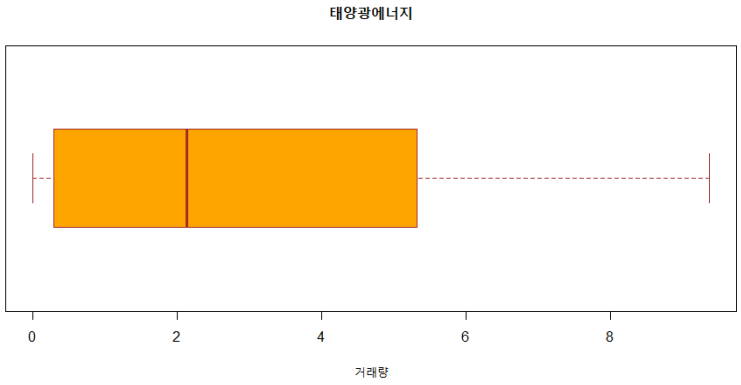


변수별로 박스플랏을 그려보면 대부분 이상치들이 상한값 쪽으로 분포되어 있음을 확인할 수 있다. 오염도가 높게 측정된 이유에는 시기별 요인과 지역별 요인 등이 있을 것이라고 예상했고, 지역별 요인은 서울시의 세분화가 큰 의미가 없을 것으로 판단하여 월별로 구분 지어 분석해보기로 하였다.

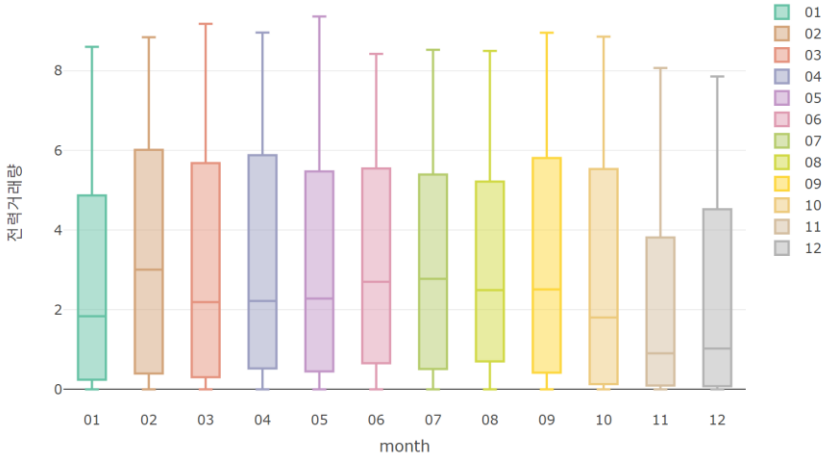


미세먼지와 이산화탄소 농도의 월별 측정값을 활용한 박스플랏이다. 두 그래프를 비교해보면 어느정도 분포 폭의 차이는 있지만 월별 평균의 흐름이 비슷한 패턴임을 알 수 있고, 이를 통해 측정값이 시기별 특성의 영향을 받는 것을 예측해볼 수 있다. 따라서, 월별 오염도의 합을 활용하여 에너지 효율과의 상관관계를 확인해 보기로 하였다.

<에너지 데이터>



특별한 이상치는 확인되지 않지만 사분위값을 기준으로 평균과 분포가 왼쪽으로 치우쳐 진 것을 확인할 수 있다. 이는 데이터 프레임을 확인해보면 시간에 따라 전력의 양 차이가 크지만 높게 측정되는 시간이 낮 1~3시 사이로 한정적이기 때문임을 예측해볼 수 있다. 하지만 이러한 차이가 매일 반복되는 패턴을 가지고 있기 때문에 이상치가 존재하지 않고, 에너지 데이터 또한 시기별 측정값의 합산을 상관 분석에 활용해보아도 문제가 없다고 판단하였다.

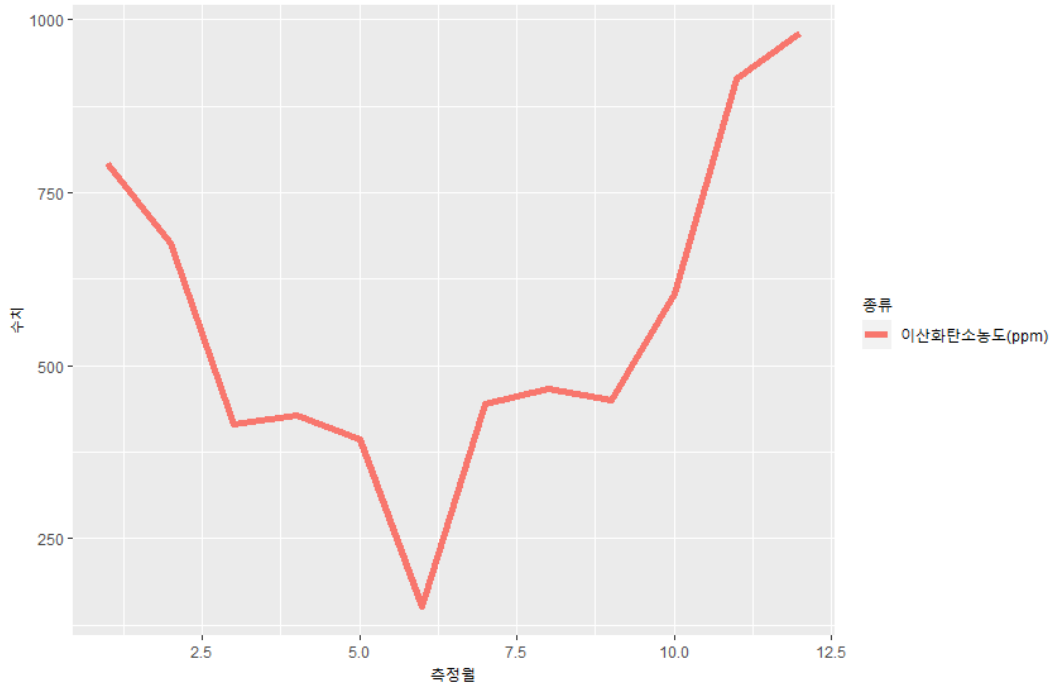


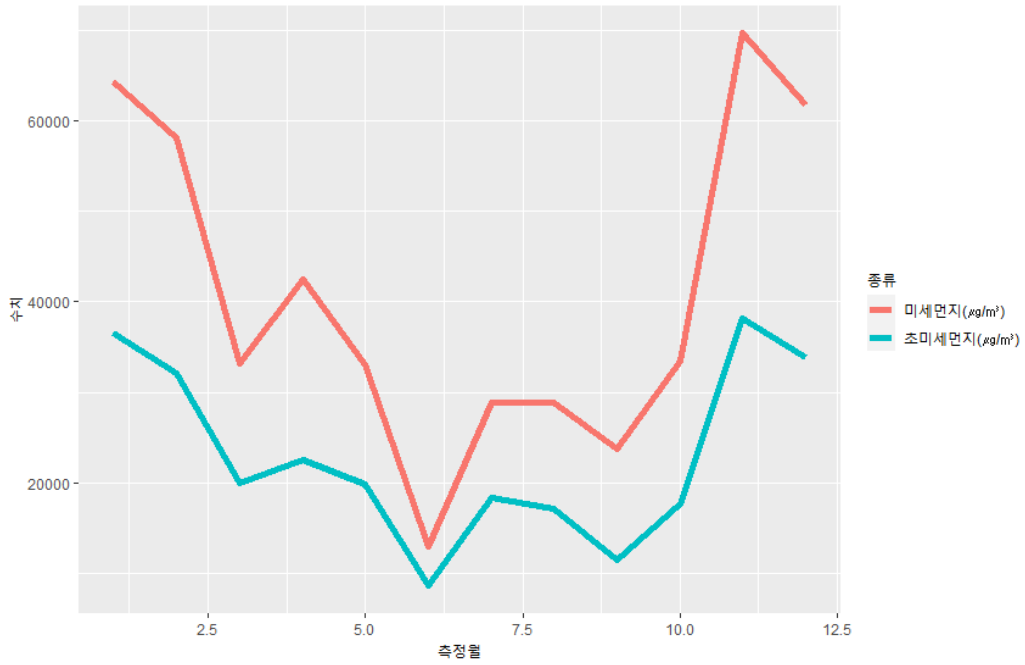
2.3 데이터 분석 및 시각화

2.3.1. 데이터 시각화

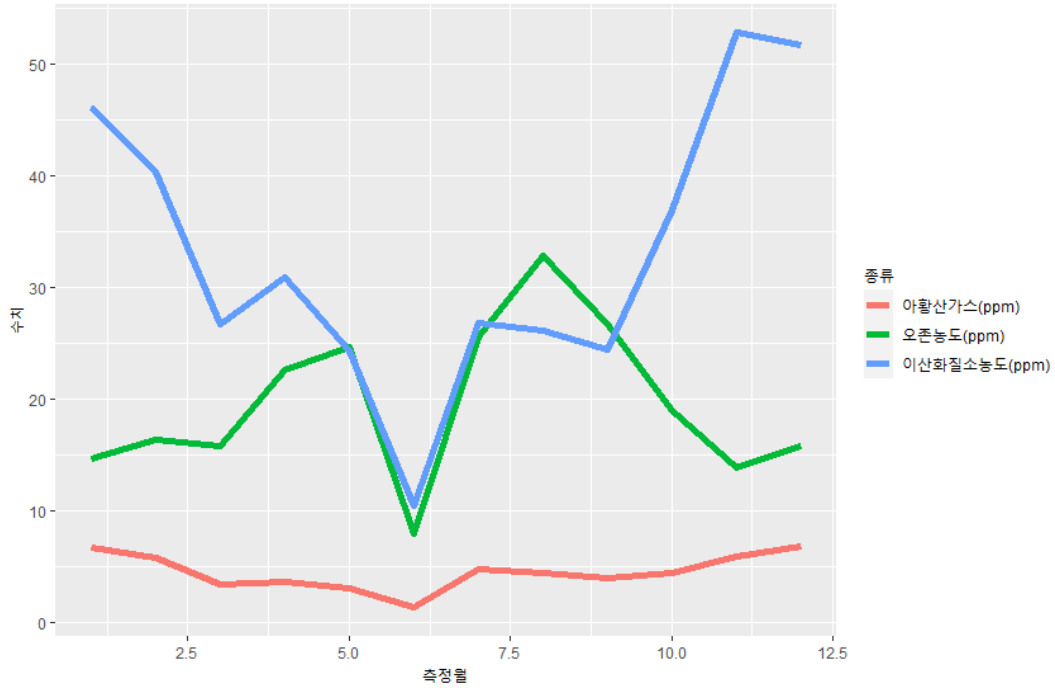
<2018년 월별 서울시 대기오염도 및 태양광 발전량>

변수별로 측정 단위의 차이가 크기 때문에 한번에 그래프로 표현하는 데에 어려움이 있어 비슷한 단위의 변수들만 한번에 그려보았다.

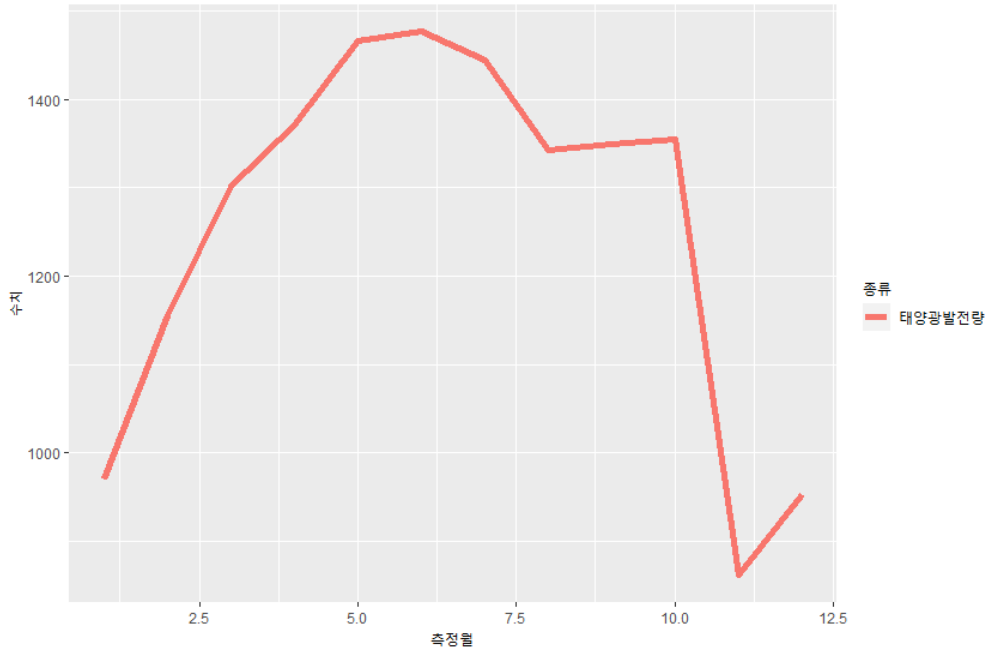




이산화탄소와 미세먼지 수치 모두 6월달에 가장 낮고 11~12월에 가장 낮게 측정된 것을 확인할 수 있다. 세 종류의 오염도 수치는 비슷한 흐름을 보인다.



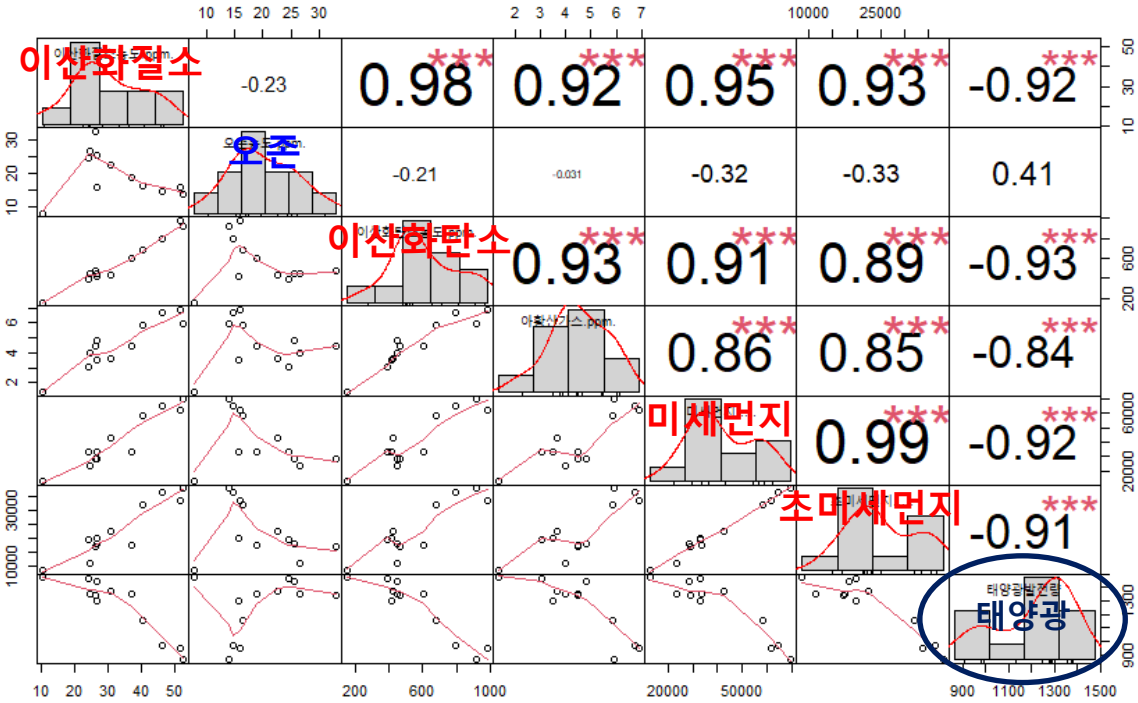
이산화질소는 앞서 보았던 이산화탄소와 미세먼지와 같은 흐름을 보이고 있지만, 오존농도와 아황산가스는 조금 다른 패턴을 보인다. 아황산가스의 경우 기존 측정 수치가 매우 낮은 단위이기 때문에 월별 수치의 차이가 크지 않은 것으로 나타나는데, 태양광 전력과의 상관관계는 이어서 상관분석을 통해 조금 더 자세한 분석이 필요할 것으로 생각된다.



태양광 전력의 경우 오염도 측정량과 반대의 흐름을 보이고 있다. 미세먼지나 이산화탄소 등의 수치가 가장 높았던 6월달에 태양광 전력의 수치가 가장 낮게 측정되었고, 12월에 가장 낮은 수치를 보여주고 있다.

2.3.2. 상관관계 분석 및 시각화

이산화질소농도, 오존농도, 이산화탄소농도, 아황산가스, 미세먼지, 초미세먼지, 태양광 발전량 변수를 이용하여 상관계수를 확인하였다. 상관관계 결과에 따르면 오존 농도를 제외한 오염도 변수들은 태양광 변수와 강한 음의 선형관계를 가지고 있다.



<회귀분석>

태양광을 종속변수로, 나머지 오염도 측정 변수들을 독립변수로 하여 회귀분석을 시도하였다. Full model의 경우 상관성이 높은 독립변수간의 상관성이 높아 다중공선성의 문제가 있기 때문에 단계별 변수선택법을 활용하여 적합한 모델을 찾기로 하였다.

단계별 선택법(Stepwise)

```
Call:
lm(formula = st_태양광 ~ st_오존 + st_이산화탄소, data = st_mdata1)
```

단계별 선택법의 결과 오존과 이산화탄소의 농도가 포함된 모형이 선택되었다. 이산화탄소, 이산화질소, 미세먼지 등의 독립변수들이 강한 양의 상관성을 갖기 때문에 하나의 변수만 선택되고, 다른 방향의 설명력을 가진 오존이 또 다른 독립변수로 선정된 것으로 예상된다.

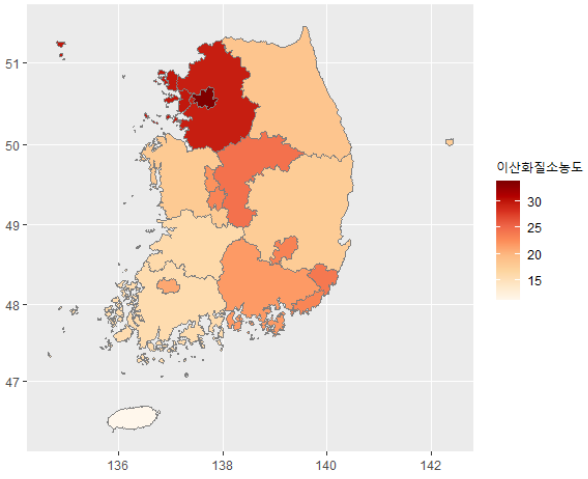
모형 : 태양광 ~ 이산화탄소 + 오존

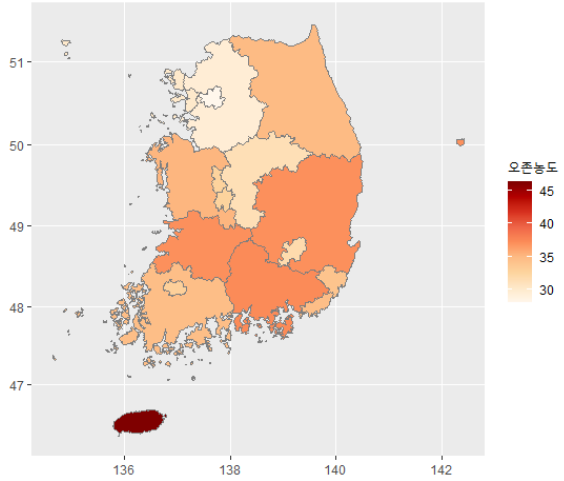
```
Coefficients:
                              Estimate                Std. Error t value Pr(>|t|)
(Intercept)     0.00000000000000004631  0.10464533931745124973    0.000   1.0000
st_이산화탄소 -0.83866917559134568361  0.11261652606797561971   -7.447 0.000039 ***
st_오존          0.27745640300740104234  0.11261652606797563358    2.464   0.0359 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3625 on 9 degrees of freedom
Multiple R-squared:  0.8925,    Adjusted R-squared:  0.8686
F-statistic: 37.35 on 2 and 9 DF,  p-value: 0.00004381
```

모형 적합 결과 결정계수 0.8925이고, F통계량의 p-value가 0.00004로 유의수준 5%하에서 매우 유의하다. 따라서 해당 모형은 자료를 잘 설명하고 있음을 알 수 있다. 독립변수들의 p-value또한 모두 0.05보다 작으므로 유의미한 설명력을 가지고 있고, 이산화탄소가 적어지고 오존 농도가 높아질수록 태양광에너지 전력이 높아진다고 할 수 있다.

2.4 데이터 분석 결과 –태양광 생산에 효율적인 지역 선정

name	시도별(2)	2018. 01	2018. 02	2018. 03	2018. 04	2018. 05	2018. 06	2018. 07	2018. 08	2018. 09	2018. 10	2018. 11	2018. 12	이산화질소농도
서울특별시	서울	0.035	0.034	0.033	0.031	0.025	0.024	0.019	0.017	0.019	0.028	0.039	0.035	33.9
부산광역시	부산	0.022	0.022	0.021	0.019	0.018	0.019	0.016	0.012	0.014	0.019	0.027	0.023	23.2
대구광역시	대구	0.027	0.025	0.022	0.017	0.013	0.013	0.01	0.01	0.014	0.021	0.033	0.029	23.4
인천광역시	인천	0.029	0.027	0.029	0.027	0.023	0.022	0.017	0.016	0.019	0.026	0.035	0.03	30
광주광역시	광주	0.024	0.024	0.019	0.016	0.013	0.011	0.009	0.008	0.013	0.019	0.03	0.024	21
대전광역시	대전	0.026	0.028	0.022	0.017	0.014	0.014	0.011	0.011	0.012	0.02	0.03	0.028	23.3
울산광역시	울산	0.024	0.024	0.023	0.021	0.019	0.02	0.017	0.013	0.013	0.018	0.027	0.022	24.1
세종특별자치시	세종	0.028	0.025	0.019	0.015	0.013	0.014	0.011	0.008	0.012	0.021	0.031	0.026	22.3
경기도	도평균	0.033	0.031	0.028	0.025	0.021	0.019	0.015	0.014	0.017	0.025	0.036	0.032	29.6
강원도	도평균	0.023	0.02	0.018	0.014	0.013	0.012	0.01	0.01	0.01	0.015	0.022	0.02	18.7
충청북도	도평균	0.029	0.025	0.02	0.017	0.014	0.015	0.016	0.016	0.017	0.021	0.029	0.027	24.6
충청남도	도평균	0.022	0.02	0.016	0.014	0.011	0.01	0.008	0.009	0.011	0.016	0.023	0.021	18.1
전라북도	도평균	0.019	0.018	0.016	0.014	0.011	0.01	0.008	0.007	0.009	0.013	0.019	0.017	16.1
전라남도	도평균	0.016	0.017	0.015	0.013	0.011	0.011	0.008	0.008	0.01	0.013	0.02	0.016	15.8
경상북도	도평균	0.018	0.019	0.016	0.013	0.012	0.011	0.01	0.01	0.012	0.015	0.022	0.02	17.8
경상남도	도평균	0.022	0.023	0.02	0.018	0.016	0.016	0.013	0.012	0.012	0.017	0.027	0.022	21.8
제주특별자치도	도평균	0.011	0.012	0.012	0.01	0.009	0.008	0.006	0.006	0.007	0.009	0.013	0.01	11.3





회귀분석 결과에 따라 태양광 에너지 생산에 적합한 지역을 탐색하기 위해 시각화를 시도하였다. 유의한 변수로 선정된 이산화탄소 농도의 2018년 전국 데이터를 구하지 못하여, 가장 비슷한 설명력을 가진 이산화질소의 도시별 평균값을 이용하였다.
그 결과 서울의 이산화질소 수치는 33.9로 가장 높았고, 제주도가 11.3으로 가장 낮았다. 그 외에 전라도와 경상북도, 충청남도가 낮은 수치를 보이고 있다. 오존농도는 제주도가 46.5로 가장 높고 서울이 28로 가장 낮았다. 그 외에는 경상도와 전라북도가 높은 수치를 보이고 있다. 이러한 수치를 통해서 낮은 이산화질소 수치와 높은 오존 농도를 가진 제주도 또는 경북, 전북 지역이 태양광 에너지를 생산하는 데에 타 지역들보다 효율적이라는 사실을 확인할 수 있다.

3. 기대 효과

정부의 탈원전을 통한 친환경 에너지를 지향함에 있어 효율적인 지역선정 하는데 가볍게 참고 할 수 있는 내용이 될 것이라 생각한다. 하지만 대기오염 수치는 지역 특성이나 외부요인의 영향을 받기 쉬운 수치이며 지형에 따른 풍량과 태양열 노출지수 같은 여러가지 변수를 생각해야 하기 때문에 단순하게 몇가지의 변수를 통해 지역선정을 하는데 단정지을 가벼운 문제는 아니다.

하지만 친환경 에너지라는 명목상으로 무분별한 세금지원으로 수도권의 태양광 시설을 구축하기보다, 데이터가 보여주는 객관적인 통계를 활용하여 정부의 정책에 효율성을 높이고자 시도하였으며, 유의미한 결과가 나왔다고 생각한다.