데이터분석기초 팀 프로젝트

팀 152. 명율합작

성명	학번	담당업무
이재욱	2016310179	데이터 수집 및 코드 작성
이환	2021314833	코드 작성 및 보고서 작성
정희진	2021311469	발표 자료 제작
최준영	2021312353	발표 영상 촬영

1. 개요

*** 프로젝트 주제 및 분석 목적**

주제: 서울 미세먼지 수치와 서울대공원 입장객수의 상관관계

가설: 미세먼지 수치가 낮은 날일수록 놀이공원의 이용객 수가 높을 것이다. (사람들이 야외 놀이공원을 이용함에 있어 미세먼지 수치를 고려할 것이다)

*** 데이터 수집 방법**

D(Data): 서울시 도로변 기간별 일평균 대기환경 현황/서울대공원 일별 입장객 현황

서울/수도권 미세먼지 일일 데이터

http://data.seoul.go.kr/dataList/OA-2224/S/1/datasetView.do;jsessionid=F372CF813E70ADD988AAE03D3DEE21C7.new_portal-svr-21

서울대공원 일일 입장객 데이터

http://data.seoul.go.kr/dataList/OA-15386/F/1/datasetView.do

2. 데이터 분석 과정

분석의 용이를 위해 2022년 1월 1일부터 10월 31일까지의 기간으로 범위를 한정하고, 미세먼지 데이터의 경우 여러 측정소의 데이터 중 종로 측정소의 데이터를 대표 기준으로 설정하였다.

*** EDA**

미세먼지 데이터의 탐색적 데이터 분석 결과는 다음과 같다. 총 데이터 수는 304개로, 확인해본 결과, 결측치는 없었다. 평균값은 소수점아래 둘째 자리에서 반올림하였을 때 약 36.58(㎍/㎥)로 나타났다.

In [82]:
```
import pandas as pd
```

In [83]:
```
# 20220101-20221031 까지의 종로 측정소의 미세먼지 데이터
dust_data = pd.read_csv('종로미세먼지_1_10.csv', encoding = 'cp949')
dust_data.head()
```

Out [83]:

	측정일자	도로변구분	측정소명	미세먼지(㎍/㎥)	오존(ppm)	이산화질소농도(ppm)	일산화탄소농도(ppm)	아황산가스농도(ppm)
0	20220101	일반도로	종로	24	0.021	0.025	0.5	0.003
1	20220102	일반도로	종로	22	0.025	0.020	0.4	0.003
2	20220103	일반도로	종로	22	0.012	0.023	0.5	0.003
3	20220104	일반도로	종로	20	0.021	0.014	0.4	0.003
4	20220105	일반도로	종로	20	0.026	0.015	0.4	0.002

In [84]:
```
# 데이터 정보
dust_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 304 entries, 0 to 303
Data columns (total 8 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   측정일자          304 non-null    int64
 1   도로변구분         304 non-null    object
 2   측정소명          304 non-null    object
 3   미세먼지(㎍/㎥)     304 non-null    int64
 4   오존(ppm)        304 non-null    float64
 5   이산화질소농도(ppm)  304 non-null    float64
 6   일산화탄소농도(ppm)  304 non-null    float64
 7   아황산가스농도(ppm)  304 non-null    float64
dtypes: float64(4), int64(2), object(2)
memory usage: 19.1+ KB
```

In [85]:
```
# 결측치 검사
dust_data.isnull().sum()
```

Out [85]:
```
측정일자            0
도로변구분           0
측정소명            0
미세먼지(㎍/㎥)       0
오존(ppm)          0
이산화질소농도(ppm)    0
일산화탄소농도(ppm)    0
아황산가스농도(ppm)    0
dtype: int64
```

In [86]:
```
# 미세먼지에 대한 통계량 도출
micro_data = dust_data.loc[:, '미세먼지(㎍/㎥)']
micro_data.describe()
```

Out [86]:
```
count    304.000000
mean      36.582237
std       20.051579
min        6.000000
25%       24.000000
50%       32.000000
75%       44.000000
max      139.000000
Name: 미세먼지(㎍/㎥), dtype: float64
```

서울대공원 입장객 데이터 역시 같은 기간으로 한정하였으므로 데이터의 개수는 304개이며, 결측치는 없었다. 평균값의 경우 놀이공원의 특성을 고려하여 주중 및 주말 데이터를 분리하여 처리할 것이기에 추후 도출하였다. EDA 결과는 다음과 같다.

In [87]:
```
# 20220101~20221031 까지의 서울대공원 입장객 데이터
entrance_data=pd.read_csv("서울대공원 입장객_1_10.csv", encoding='UTF-8')
entrance_data.head()
```

Out[87]:

	날짜	요일	유료일계	무료일계	일합계
0	20220101	토	831	212	1043
1	20220102	일	780	304	1084
2	20220103	월	198	126	324
3	20220104	화	177	240	417
4	20220105	수	166	183	349

In [88]:
```
# 데이터 정보
entrance_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 304 entries, 0 to 303
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   날짜      304 non-null    int64 
 1   요일      304 non-null    object
 2   유료일계    304 non-null    int64 
 3   무료일계    304 non-null    int64 
 4   일합계     304 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 12.0+ KB
```

In [89]:
```
# 결측치 검사
entrance_data.isnull().sum()
```

Out[89]:
```
날짜      0
요일      0
유료일계    0
무료일계    0
일합계     0
dtype: int64
```

*** 데이터 병합 과정**

두 데이터를 병합하여 주요 변수 및 활용할 지표들인 미세먼지 측정일자, 요일, 미세먼지, 해당일 일일 입장객으로 구성된 데이터프레임을 만들었다.

In [90]:
```
# 데이터 병합 과정
day_list=list(entrance_data["요일"])
dailytotal_list=list(entrance_data["일합계"])
```

In [91]:
```
dust_data["요일"]=day_list
dust_data["공원 일일 입장객"]=dailytotal_list
```

In [92]:
```
dust_data.head()
```

Out[92]:

	측정일자	도로변구분	측정소명	미세먼지(㎍/㎥)	오존(ppm)	이산화질소농도(ppm)	일산화탄소농도(ppm)	아황산가스농도(ppm)	요일	공원 일일 입장객
0	20220101	일반도로	종로	24	0.021	0.025	0.5	0.003	토	1043
1	20220102	일반도로	종로	22	0.025	0.020	0.4	0.003	일	1084
2	20220103	일반도로	종로	22	0.012	0.023	0.5	0.003	월	324
3	20220104	일반도로	종로	20	0.021	0.014	0.4	0.003	화	417
4	20220105	일반도로	종로	20	0.026	0.015	0.4	0.002	수	349

In [93]:
```
main_data = dust_data[["측정일자", "요일", "미세먼지(㎍/㎥)", "공원 일일 입장객"]]
main_data.head()
```

Out[93]:

	측정일자	요일	미세먼지(㎍/㎥)	공원 일일 입장객
0	20220101	토	24	1043
1	20220102	일	22	1084
2	20220103	월	22	324
3	20220104	화	20	417
4	20220105	수	20	349

분석에 있어, 놀이공원의 특성상 주중과 주말의 이용객에는 유의미한 차이가 날 수밖에 없기에, 주중과 주말 데이터를 분리하여 각각을 분석하는 것이 합리적이라고 판단하여 이러한 방식으로 분석을 진행하였다.

*** 데이터 분석 과정**

주중의 경우 평균은 올림을 해서 약 2705명으로 나타났으며, 총 데이터는 216개로 나타났다.

공휴일의 경우 미세먼지가 아닌 외부요인 중 꽤 큰 영향을 줄 수 있는 요소라 판단되어, 공휴일의 방문객 수 중 Q3+1.5*IQR(주중: 7598.125, 주말: 36080.625)보다 큰 경우를 이상치로 설정하고 이를 제거하였다. 놀이공원의 특성을 고려하여 큰 값만 제거하였다.

평일의 경우 대통령 선거일 0309, 어린이날 0505, 지방선거일 0601, 현충일 0606, 추석 0909, 추석 대체휴일 0912가 제거되었다.

In [94]:
```
# 주중 데이터 분석
weekday_data = main_data[main_data['요일'].str.contains('월|화|수|목|금', na = False)]
weekday_data.head()
```

Out [94]:

	측정일자	요일	미세먼지(㎍/㎥)	공원 일일 입장객
2	20220103	월	22	324
3	20220104	화	20	417
4	20220105	수	20	349
5	20220106	목	13	933
6	20220107	금	19	463

In [95]:
```
# 통계량 도출
weekday_daily = weekday_data.loc[:, '공원 일일 입장객']
weekday_daily.describe()
```

Out [95]:
```
count      216.000000
mean      2704.199074
std       3954.749482
min         55.000000
25%        737.500000
50%       1645.500000
75%       3481.750000
max      46381.000000
Name: 공원 일일 입장객, dtype: float64
```

In [96]:
```
# 특별한 날의 이상치(공휴일 및 기념일들 중 방문객 수가 Q3+1.5*IQR= 7598.125보다 큰 경우) 제거
weekday_pro = weekday_data.drop(index=[67, 124, 151, 156, 251, 254])
```

In [97]:
```
print('Data size: ', end = '')
print(weekday_pro.shape)
```

```
Data size: (210, 4)
```

이상치 제거 이후 평일의 데이터는 총 210개로, 아래는 이를 바탕으로 회귀분석을 진행한 결과이다. 시각화 부분의 경우 이후 따로 서술하였다.

- 다음 장으로

In [98]:
```
# 회귀 분석 진행
from sklearn.linear_model import LinearRegression
import matplotlib.pyplot as plt
```

In [99]:
```
# 모델 학습
weekday_model = LinearRegression()

X1 = weekday_pro['미세먼지(㎍/㎥)']
y1 = weekday_pro['공원 일일 입장객']

weekday_model.fit(X1.values.reshape(-1,1), y1)
```

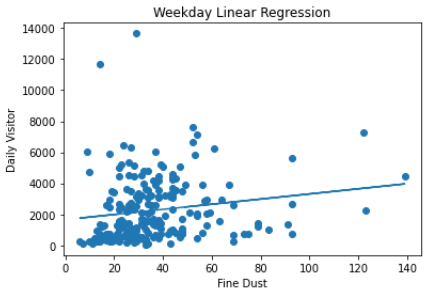
Out[99]:
```
LinearRegression()
```

In [100]:
```
# 기울기, y절편
print("coef:", weekday_model.coef_)
print("intercpt:", weekday_model.intercept_)
```

```
coef: [16.57821349]
intercpt: 1657.6393552745035
```

In [101]:
```
# 시각화
plt.title("Weekday Linear Regression")
plt.xlabel("Fine Dust")
plt.ylabel("Daily Visitor")
plt.scatter(X1, y1)
plt.plot(X1, weekday_model.coef_*X1 + weekday_model.intercept_)
```

Out[101]:
```
[<matplotlib.lines.Line2D at 0x24ae5197970>]
```



주말의 경우 따로 제거할 데이터는 없었으며, 데이터는 총 88개로, 역시 이를 바탕으로 회귀분석을 진행하였다.

In [102]:
```
# 주말 데이터 분석
weekend_data = main_data[main_data['요일'].str.contains('토|일', na = False)]
weekend_data.head()
```

Out[102]:

	측정일자	요일	미세먼지(㎍/㎥)	공원 일일 입장객
0	20220101	토	24	1043
1	20220102	일	22	1084
7	20220108	토	28	1431
8	20220109	일	28	1116
14	20220115	토	18	1433

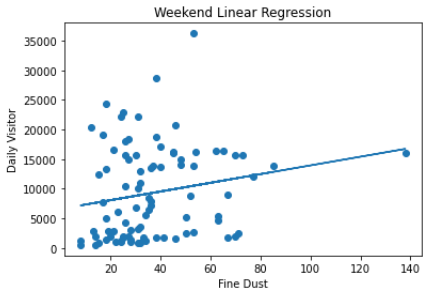
In [103]:
```
# 통계량 도출
weekend_daily = weekend_data.loc[:, '공원 일일 입장객']
weekend_daily.describe()
```

Out[103]:
```
count       88.000000
mean      9311.204545
std       7820.580316
min        504.000000
25%       1950.000000
50%       7480.000000
75%      15602.250000
max      36225.000000
Name: 공원 일일 입장객, dtype: float64
```

In [104]:
```
print('Data size: ', end = '')
print(weekend_data.shape)
```

```
Data size: (88, 4)
```

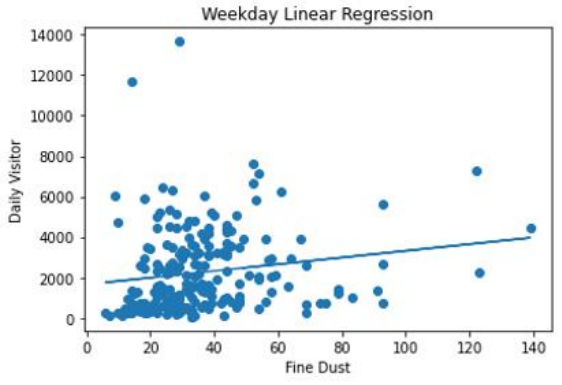
```
In [105]:
# 모델 학습
weekend_model = LinearRegression()

X2 = weekend_data['미세먼지(㎍/㎥)']
y2 = weekend_data['공원 일일 입장객']

weekend_model.fit(X2.values.reshape(-1,1), y2)

Out[105]:
LinearRegression()

In [106]:
# 기울기, y절편
print("coef:", weekend_model.coef_)
print("intercpt:", weekend_model.intercept_)

coef: [73.37309193]
intercpt: 6595.566358987559

In [107]:
# 시각화
plt.title("Weekend Linear Regression")
plt.xlabel("Fine Dust")
plt.ylabel("Daily Visitor")
plt.scatter(X2, y2)
plt.plot(X2, weekend_model.coef_*X2 + weekend_model.intercept_)

Out[107]:
[<matplotlib.lines.Line2D at 0x24ae520ffd0>]
```

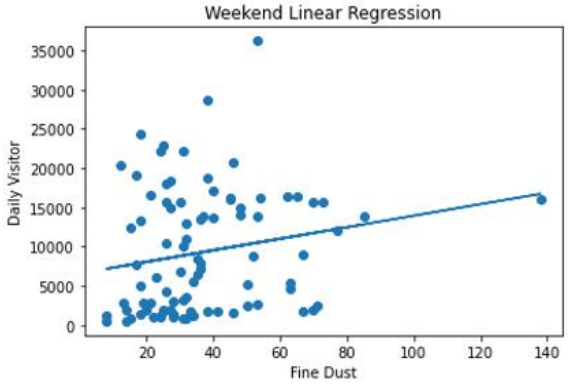


3. 분석 내용에 대한 시각화

I(Information): 두 변수 사이의 상관관계

확대된 시각화 결과는 다음 그래프와 같이 나타난다. 그러나 아래의 결과를 보았을 때 직선은 앞서 설정한 가설과 반대되는 방향으로 나타난다. 데이터의 분포를 보더라도 미세먼지와 서울대공원 방문객 수 사이에는 뚜렷한 관계가 없는 것으로 보이며, 또한 본 데이터셋에는 고미세먼지 데이터가 적어 신뢰도 또한 떨어진다고 볼 수 있을 것이다.





따라서 외부 요인을 그나마 최소화하여 추가적인 분석을 진행해보았다.

4. 보완을 위한 추가 분석 과정

이번에는 서울대공원 입장객 데이터에서 일합계 대신 유료일계를 사용하여 비교적 자발적 의사가 많이 반영된 선택이 결과에 반영되도록 하였다. 또한 앞선 결과에서 보았듯 대부분의 미세먼지 데이터가 80이하의 수치를 갖고 있으므로 해당 데이터만을 사용하여 분석을 보완해보고자 하였다.

In [108]:
```
# 새로운 기준에서의 분석을 위한 데이터프레임 재생성
n_dust_data = dust_data.copy()[["측정일자", "미세먼지(㎍/㎥)"]]
n_dust_data.head()
```

Out[108]:

	측정일자	미세먼지(㎍/㎥)
0	20220101	24
1	20220102	22
2	20220103	22
3	20220104	20
4	20220105	20

In [109]:
```
n_day_list=list(entrance_data["요일"])
n_dailytotal_list=list(entrance_data["유료일계"])
```

In [110]:
```
n_dust_data["요일"]=n_day_list
n_dust_data["일일 유료 입장객"]=n_dailytotal_list
n_dust_data.head()
```

Out[110]:

	측정일자	미세먼지(㎍/㎥)	요일	일일 유료 입장객
0	20220101	24	토	831
1	20220102	22	일	780
2	20220103	22	월	198
3	20220104	20	화	177
4	20220105	20	수	166

In [111]:
```
# 미세먼지 수치가 80보다 큰 데이터 제거
index = n_dust_data[n_dust_data['미세먼지(㎍/㎥)'] > 80].index
n_dust_data.drop(index, inplace = True)
n_dust_data
```

Out[111]:

	측정일자	미세먼지(㎍/㎥)	요일	일일 유료 입장객
0	20220101	24	토	831
1	20220102	22	일	780
2	20220103	22	월	198
3	20220104	20	화	177
4	20220105	20	수	166
...	...	...	...	...
296	20221024	31	월	2138
297	20221025	44	화	1705
301	20221029	65	토	13553
302	20221030	38	일	26574
303	20221031	49	월	2521

294 rows × 4 columns

In [112]:
```
# 주중 주말 분리 및 위와 같은 과정 반복
n_weekday_data = n_dust_data[n_dust_data['요일'].str.contains('월|화|수|목|금', na = False)]
n_weekday_data.head()
```

Out[112]:

	측정일자	미세먼지(㎍/㎥)	요일	일일 유료 입장객
2	20220103	22	월	198
3	20220104	20	화	177
4	20220105	20	수	166
5	20220106	13	목	712
6	20220107	19	금	304

이후는 처음 진행하였던 방식과 동일하게 주중과 주말을 분리하여 각각에 대한 회귀분석을 진행하였다.

In [113]:
```
# 제거하는 공휴일 데이터의 경우 유/무료와 무관하다고 판단하여 삭제 기준은 그대로 유지함
n_weekday_pro = n_weekday_data.drop(index=[67, 124, 151, 156, 251, 254])
```

In [114]:
```
n_weekday_model = LinearRegression()

X3 = n_weekday_pro['미세먼지(㎍/㎥)']
y3 = n_weekday_pro['일일 유료 입장객']

n_weekday_model.fit(X3.values.reshape(-1,1), y3)
```

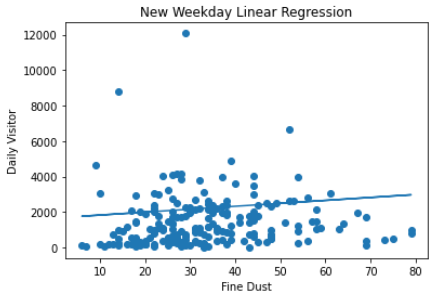
Out[114]:
```
LinearRegression()
```

In [115]:
```
print("coef:", n_weekday_model.coef_)
print("intercpt:", n_weekday_model.intercept_)
```

```
coef: [5.97450191]
intercpt: 1255.6314138269613
```

In [116]:
```
plt.title("New Weekday Linear Regression")
plt.xlabel("Fine Dust")
plt.ylabel("Daily Visitor")
plt.scatter(X3, y3)
plt.plot(X3, weekday_model.coef_*X3 + weekday_model.intercept_)
```

Out[116]:
```
[<matplotlib.lines.Line2D at 0x24ae5286dc0>]
```



In [117]:
```
n_weekend_data = n_dust_data[n_dust_data['요일'].str.contains('토|일', na = False)]
n_weekend_data.head()
```

Out[117]:

	측정일자	미세먼지(㎍/㎥)	요일	일일 유료 입장객
0	20220101	24	토	831
1	20220102	22	일	780
7	20220108	28	토	1162
8	20220109	28	일	911
14	20220115	18	토	1122

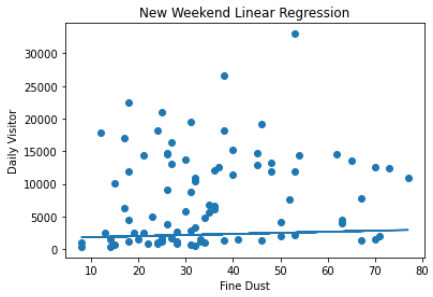
In [118]:
```
n_weekend_model = LinearRegression()

X4 = n_weekend_data['미세먼지(㎍/㎥)']
y4 = n_weekend_data['일일 유료 입장객']

n_weekend_model.fit(X4.values.reshape(-1,1), y4)
```

Out[118]:
```
LinearRegression()
```

In [119]:
```
print("coef:", n_weekend_model.coef_)
print("intercpt:", n_weekend_model.intercept_)
```

```
coef: [62.1773579]
intercpt: 5820.370885103486
```

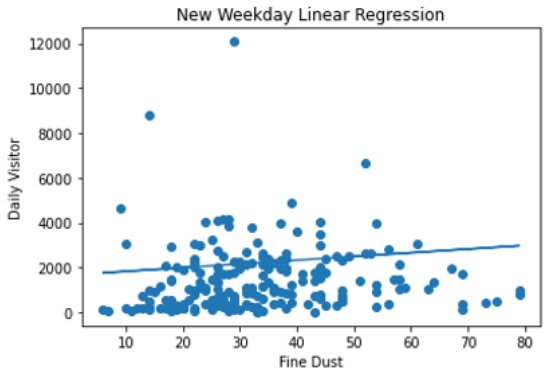
```
In [120]: plt.title("New Weekend Linear Regression")
          plt.xlabel("Fine Dust")
          plt.ylabel("Daily Visitor")
          plt.scatter(X4, y4)
          plt.plot(X4, weekday_model.coef_*X4 + weekday_model.intercept_)
Out[120]: [<matplotlib.lines.Line2D at 0x24ae52ff790>]
```

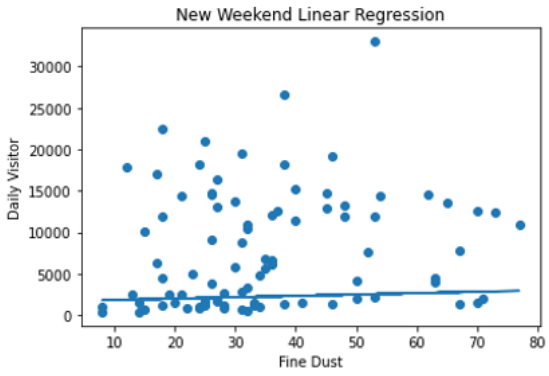



```
In [ ]:
```

K(Knowledge): 상관관계를 통해 알 수 있는 정보

조건을 변경하여 다시 한번 분석을 진행해보았지만, 둘의 상관관계는 더욱 약해지는 양상으로 나타났으며, 특히 주말의 경우 선의 기울기가 거의 0에 가까워지는 모습을 볼 수 있었다. 따라서 본 분석에서 도출된 결과에 따르면, 미세먼지 수치와 놀이공원 이용객의 수 사이에는 유의미한 상관관계가 존재하지 않는다고 볼 수 있다.





5. 결론 및 제언

W(Wisdom): 결과 분석 후 비전/해결방안 제시

앞선 분석의 결과에 따라 미세먼지 수치와 놀이공원 이용객의 수 사이에는 유의미한 상관관계가 존재하지 않는다고 볼 수 있다. 따라서 가설 '미세먼지 수치가 낮은 날일수록 놀이공원의 이용객 수가 높을 것이다.' 는 기각된다.

즉, 사람들은 장시간의 야외활동을 하는데 있어 생각보다 미세먼지를 고려하는 정도가 약하다는 것을 알 수 있다. 이러한 결과를 토대로 얻어낼 수 있는 시각은 정부의 입장과 놀이공원 운영주의 입장 두가지 정도로 정리할 수 있겠다.

우선 정부의 입장에서, 사람들이 그 유해성이 입증된 미세먼지에도 별로 개의치 않고 야외활동을 하고 있다는 점에서, 미세먼지 관련 공익광고와 캠페인에 좀 더 많은 노력을 기울여 국민의 건강을 증진하는데 도움을 줄 수 있을 것이다.

놀이공원 운영주의 입장에서는, 사람들이 미세먼지의 여부와 크게 관계없이 놀이공원을 찾게 된다는 점에서 착안하여, 미세먼지 수치가 높은 날에는 놀이공원 내에 있는 실내 시설의 활성화 혹은 실내 공간에서 진행되는 이벤트 기획 등을 진행하고, 이를 미세먼지와 관련하여 홍보함으로써, 고객의 건강을 신경 쓰는 기업이라는 긍정적 기업 이미지를 얻을 수 있을 것이다.