**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

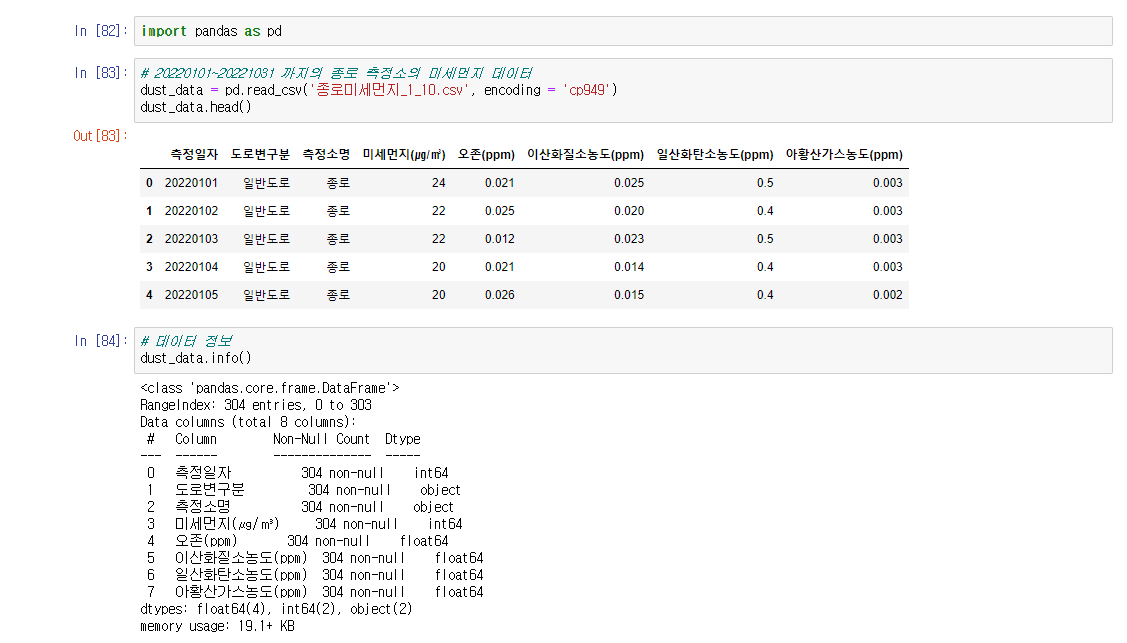
http://data.seoul.go.kr/dataList/OA-15386/F/1/datasetView.do

**2. 데이터 분석 과정**

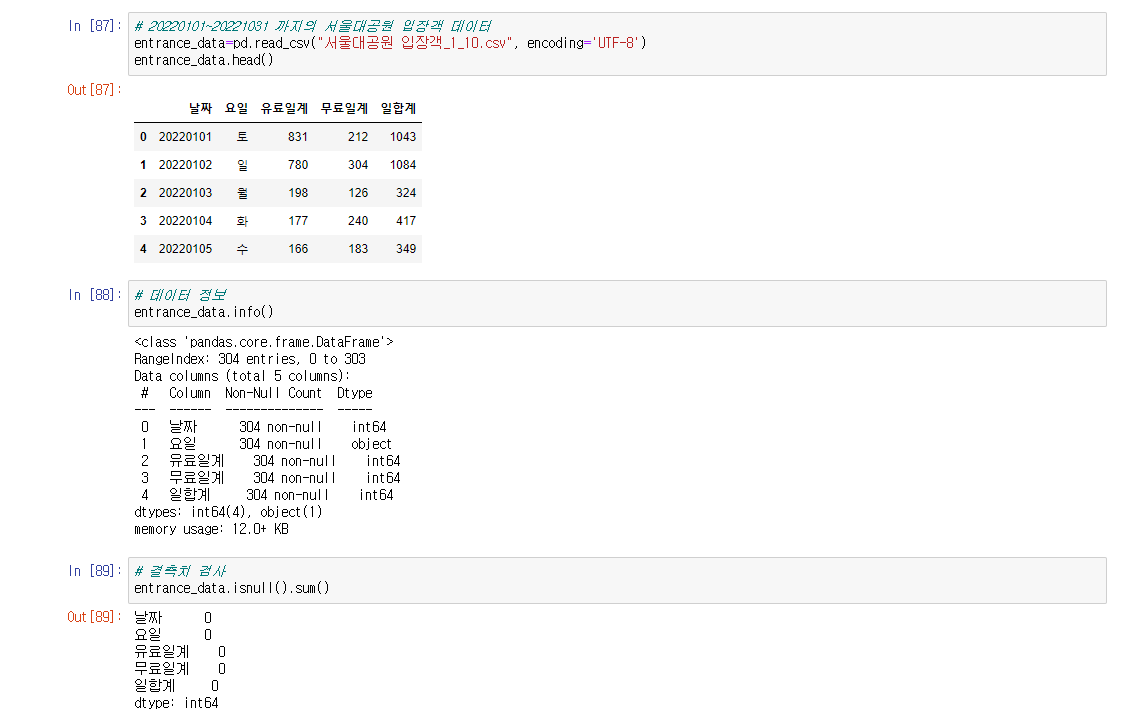
분석의 용이를 위해 2022년 1월 1일부터 10월 31일까지의 기간으로 범위를 한정하고, 미세먼지 데이터의 경우 여러 측정소의 데이터 중 종로 측정소의 데이터를 대표 기준으로 설정하였다.

**\* 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.





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.

**\* 데이터 병합 과정**

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



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.

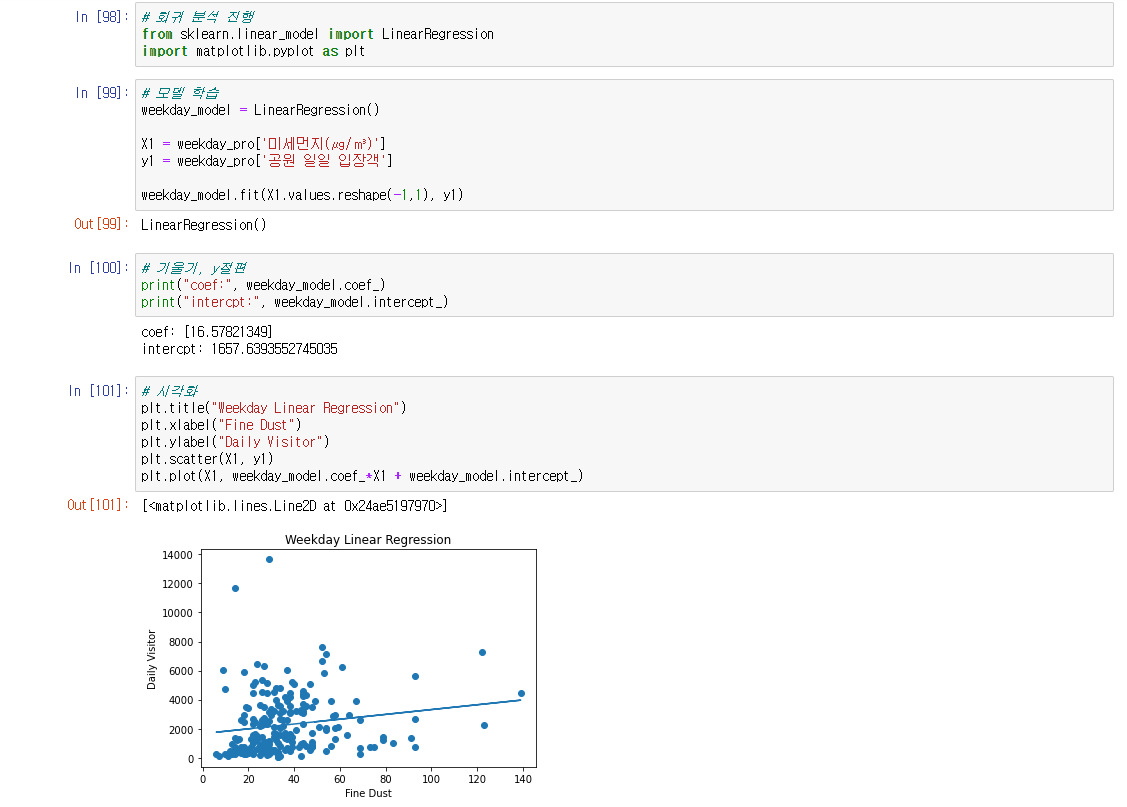
공휴일의 경우 미세먼지가 아닌 외부요인 중 꽤 큰 영향을 줄 수 있는 요소라 판단되어, 공휴일의 방문객 수 중 Q3+1.5\*IQR(주중: 7598.125, 주말: 36080.625)보다 큰 경우를 이상치로 설정하고 이를 제거하였다. 놀이공원의 특성을 고려하여 큰 값만 제거하였다.

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

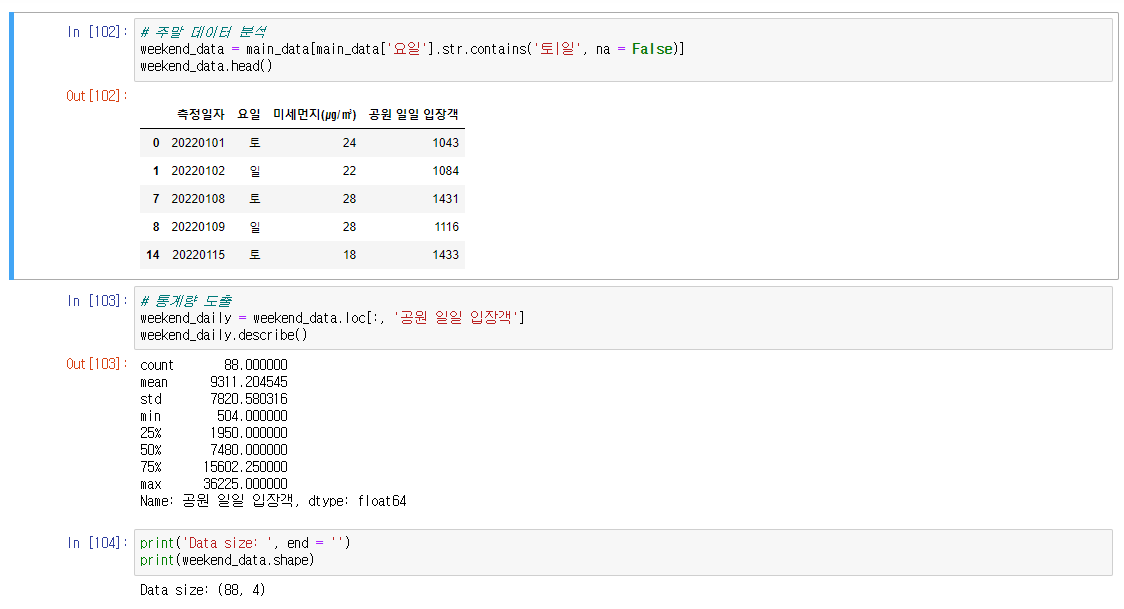


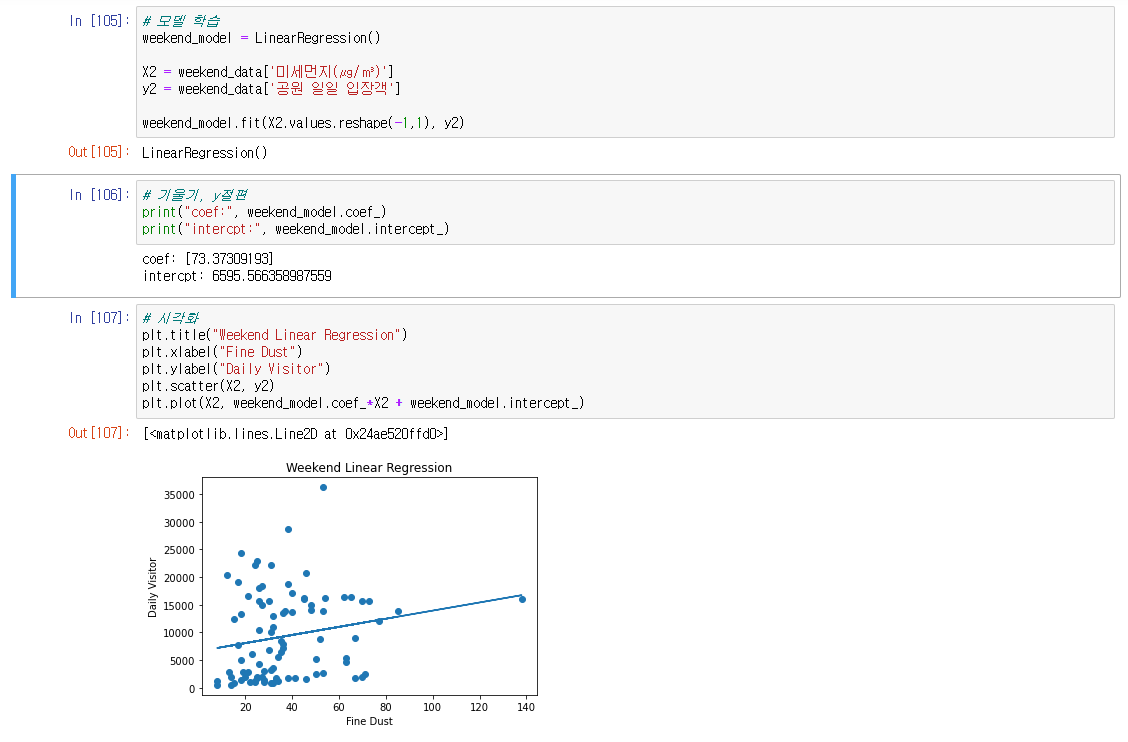
이상치 제거 이후 평일의 데이터는 총 210개로, 아래는 이를 바탕으로 회귀분석을 진행한 결과이다. 시각화 부분의 경우 이후 따로 서술하였다.

- 다음 장으로



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

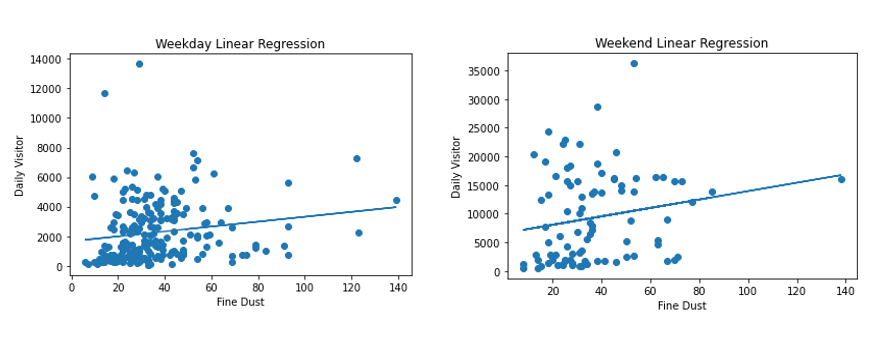




**3. 분석 내용에 대한 시각화**

**I(Information): 두 변수 사이의 상관관계**

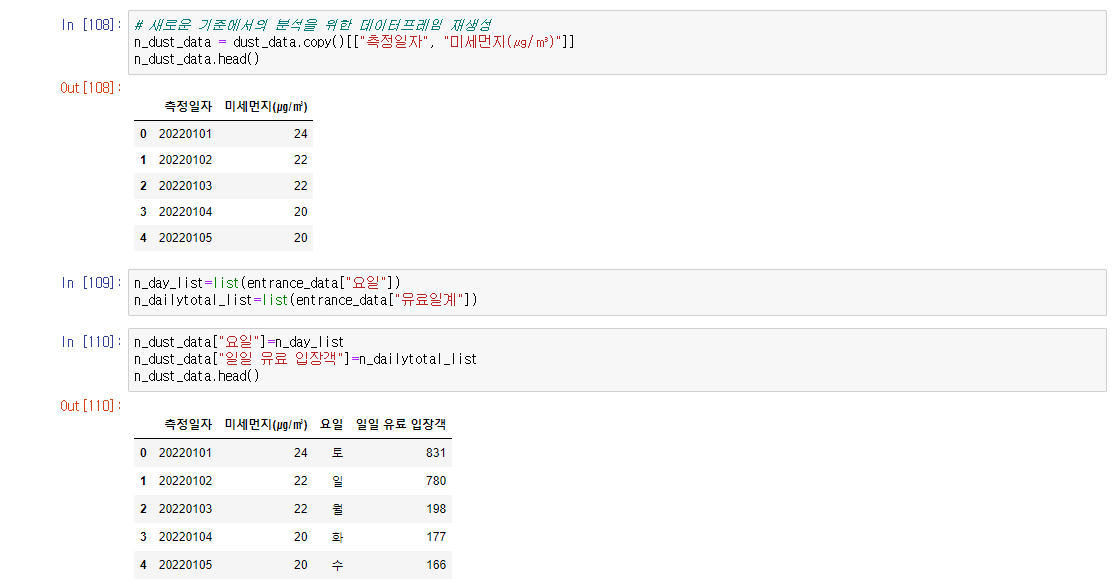
확대된 시각화 결과는 다음 그래프와 같이 나타난다. 그러나 아래의 결과를 보았을 때 직선은 앞서 설정한 가설과 반대되는 방향으로 나타난다. 데이터의 분포를 보더라도 미세먼지와 서울대공원 방문객 수 사이에는 뚜렷한 관계가 없는 것으로 보이며, 또한 본 데이터셋에는 고미세먼지 데이터가 적어 신뢰도 또한 떨어진다고 볼 수 있을 것이다.

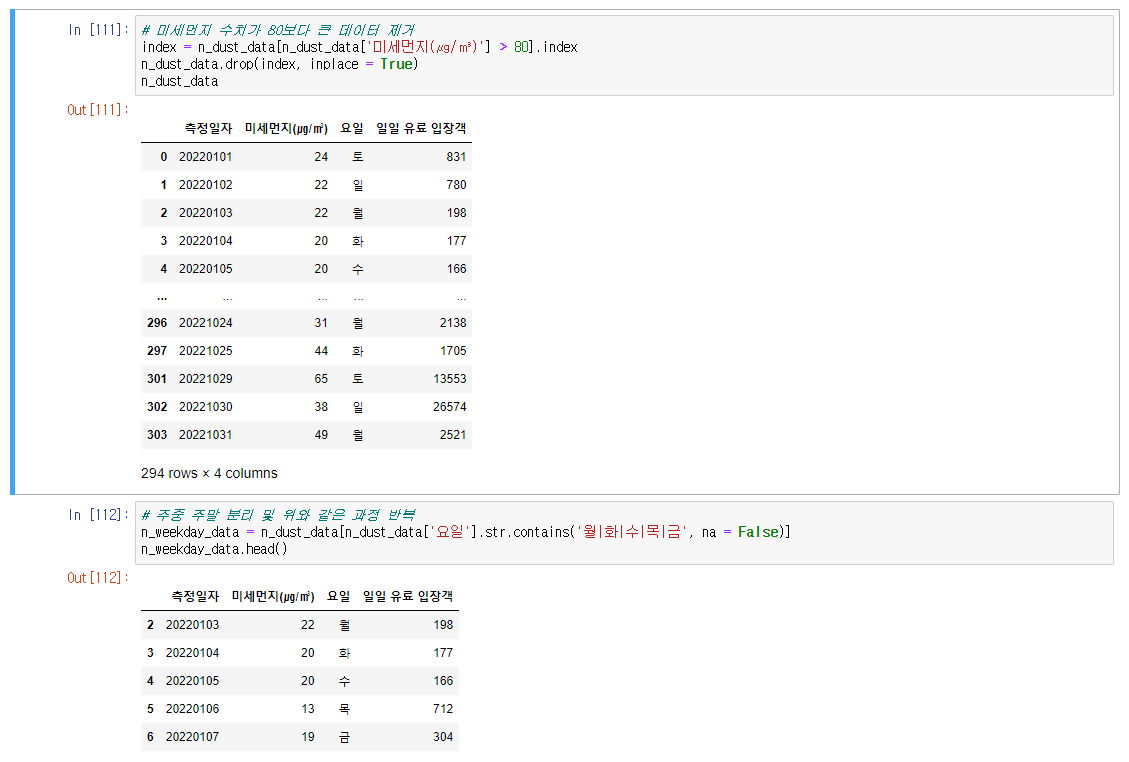


따라서 외부 요인을 그나마 최소화하여 추가적인 분석을 진행해보았다.

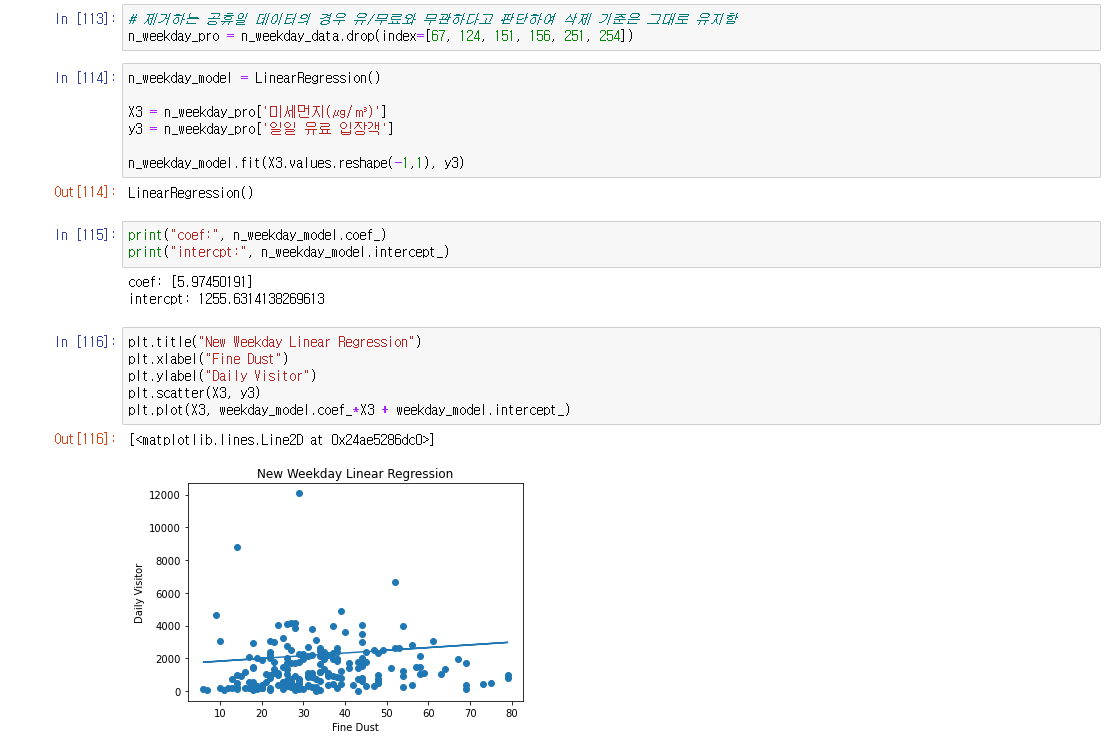
**4. 보완을 위한 추가 분석 과정**

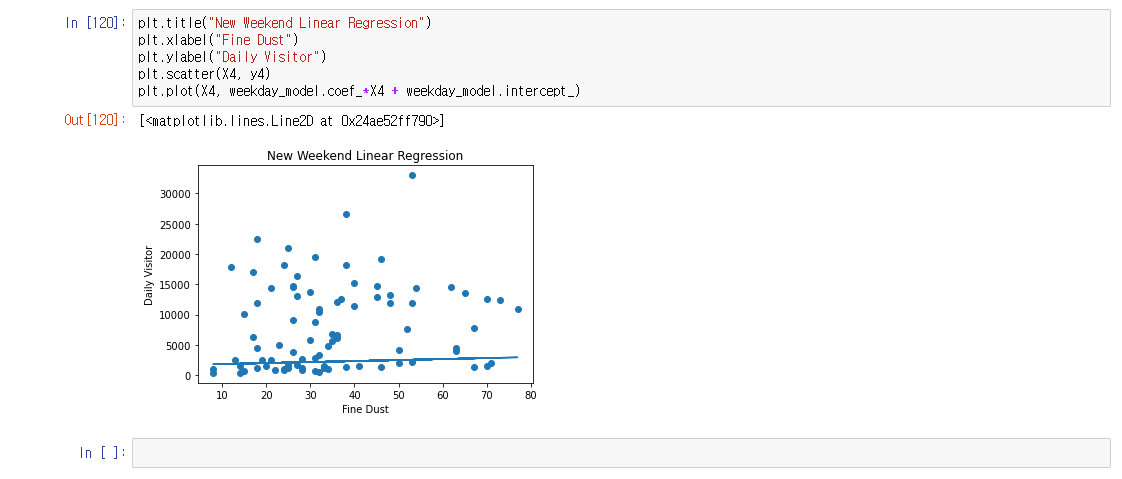
이번에는 서울대공원 입장객 데이터에서 일합계 대신 유료일계를 사용하여 비교적 자발적 의사가 많이 반영된 선택이 결과에 반영되도록 하였다. 또한 앞선 결과에서 보았듯 대부분의 미세먼지 데이터가 80이하의 수치를 갖고 있으므로 해당 데이터만을 사용하여 분석을 보완해보고자 하였다.





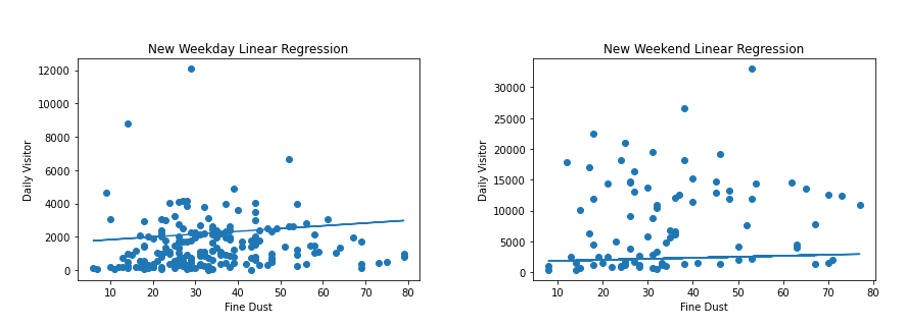
이후는 처음 진행하였던 방식과 동일하게 주중과 주말을 분리하여 각각에 대한 회귀분석을 진행하였다.





**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.