2020년 서울의 미세먼지(PM-10)

통계분석 포트폴리오

Contents

Part 1

Data Visualization

- Box-Plot
  대륙 별 미세먼지농도 수준
- Bar-Chart
  나라 별 미세먼지농도 수준

Part 2

T-Test

- One Sample
  서울의 미세먼지 수준
- Paired Sample
  정부의 미세먼지 대응책
- Independent Sample
  서울과 부산의 미세먼지

Part 3

ANOVA Test

- One-way
  계절에 따른 미세먼지

Part 4

Regression

- Simple Linear
  온도와 미세먼지I
- Polynomial
  온도와 미세먼지II
- Multiple
  온도/풍속/주말여부

미세먼지란

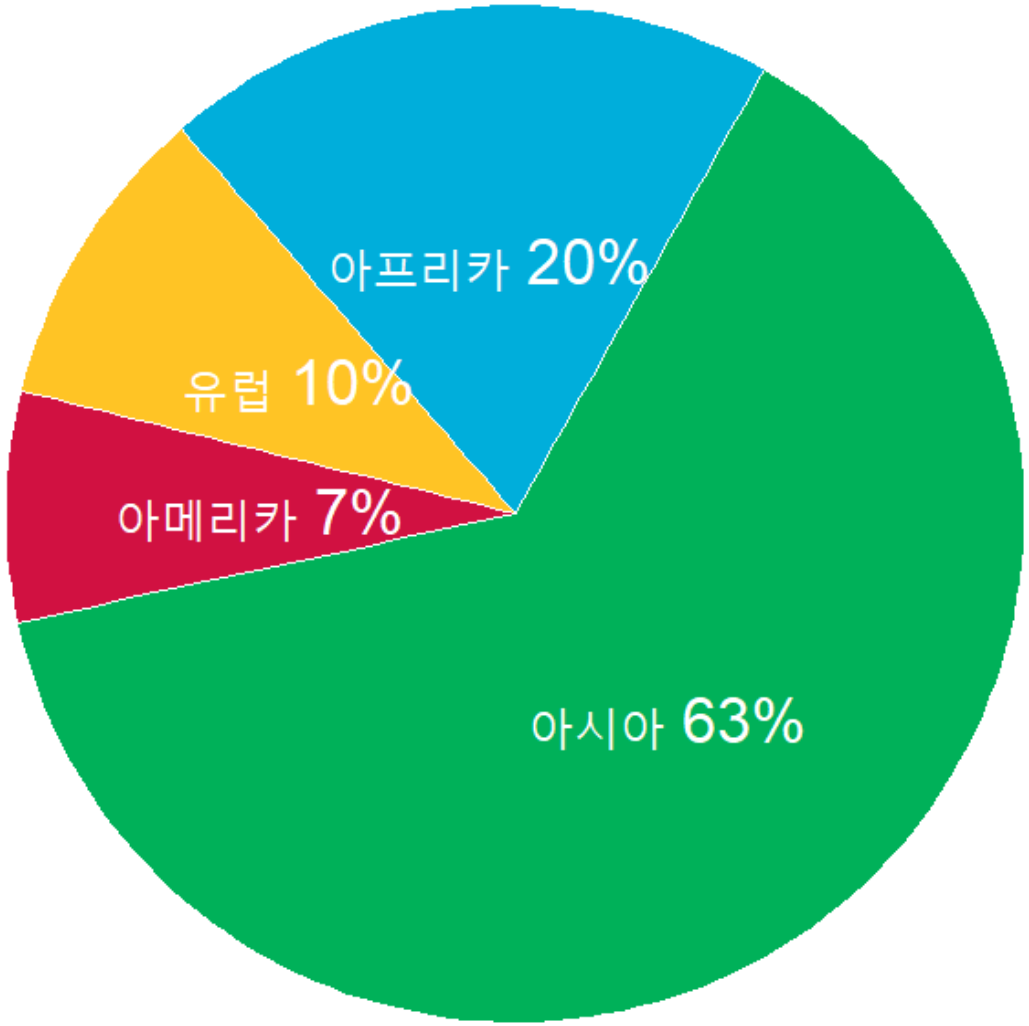
- 입자의 지름이 10 마이크로미터(㎛) 이하인 먼지(PM-10).

대륙 별 미세먼지 발생현황

- 아시아의 비율이 다른 타 대륙의 총 합보다 높다.
- 아시아 내에서도

1. 중동 : 카타르, 사우디 등..
2. 동북 : 중국,한국 등..
3. 남동 : 인도, 인도네시아 등..

으로 나뉘어 각각 3분의 1씩 미세먼지를 발생한다.



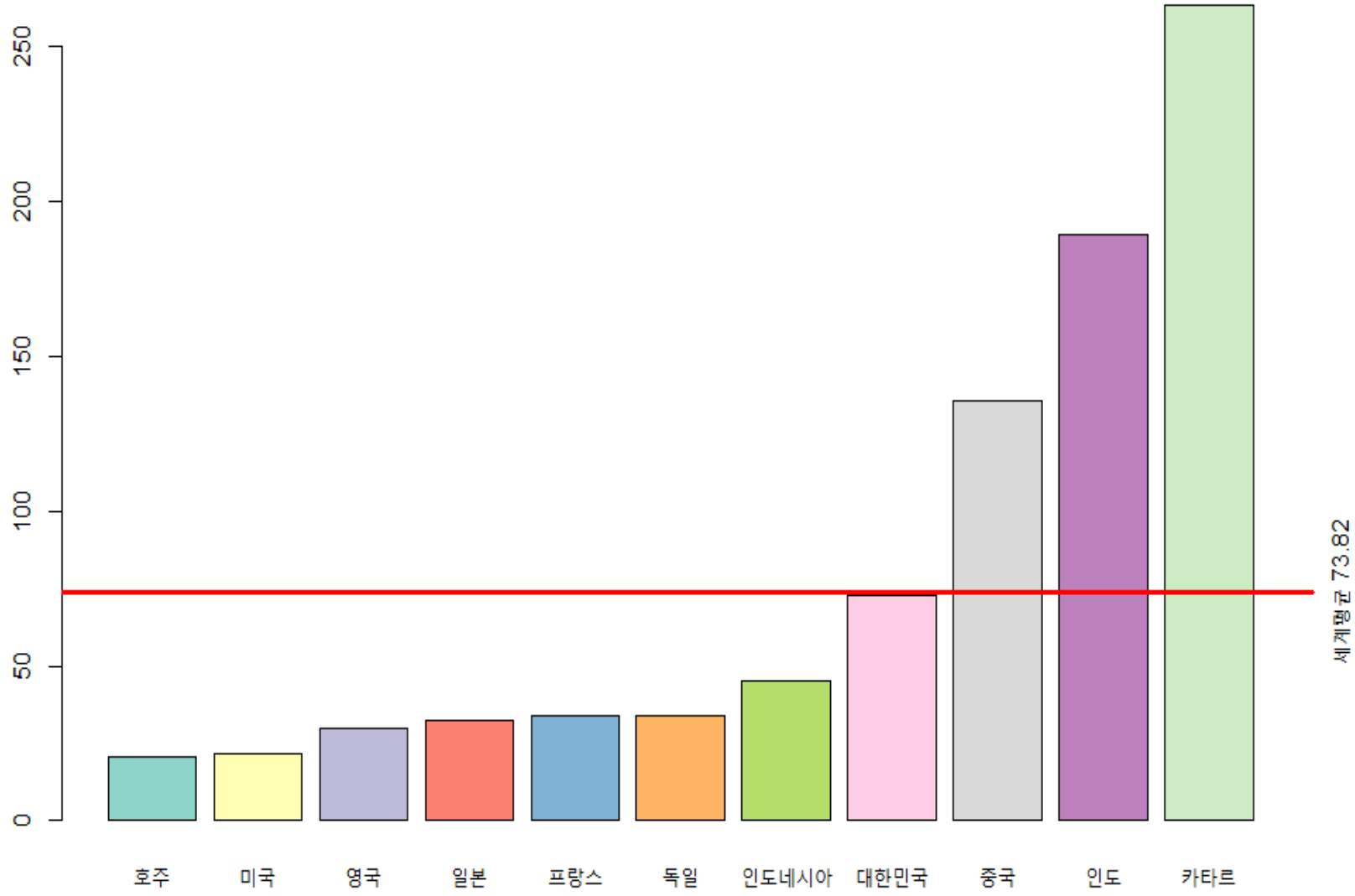
대륙 별 미세먼지 발생현황
(세계보건기구 발표 2016)

나라 별 미세먼지 발생현황

- 세계의 연평균 미세먼지 농도는 73.82 ㎍/㎥다.

- 대한민국의 연평균 미세먼지 농도는 대략 70 ㎍/㎥으로 세계평균보다 살짝 낮다.

- 가장 많은 미세먼지 발생 국가는 중동국가인 카타르로 연평균 250 ㎍/㎥을 상회한다.



나라 별 미세먼지 발생현황

(세계보건기구 발표 2016)

• 문제정의:

세계 보건기구 WHO에서 발표한 하루 미세먼지 (PM-10) 농도의 권장량은 50 ㎍/㎥ 이다.
서울의 미세먼지 농도가 가이드라인에 적합한지 알아보려 한다.

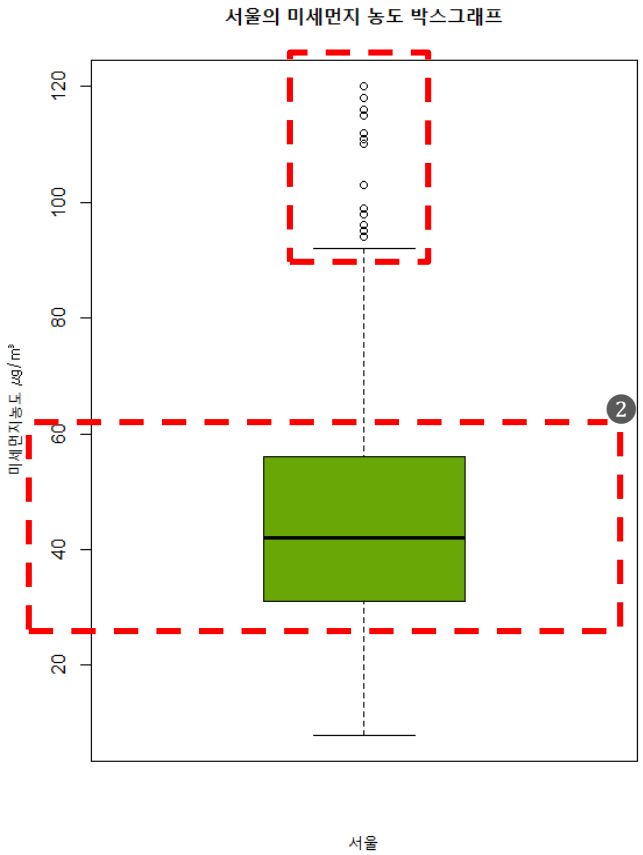
• 데이터 수집:

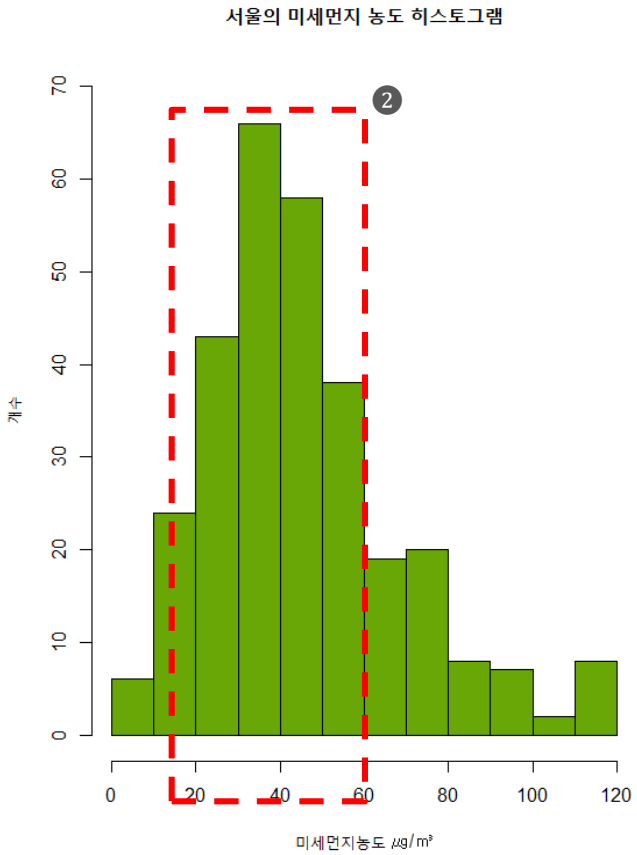
2020년 서울의 매일 평균 미세먼지 농도 측정 데이터 / 출처: 기상청

	지역	날짜	미세먼지농도
1	서울	2020-01-01	39
2	서울	2020-01-02	69
3	서울	2020-01-03	78
4	서울	2020-01-04	63
5	서울	2020-01-05	62
6	서울	2020-01-06	46
7	서울	2020-01-07	10
8	서울	2020-01-08	32
9	서울	2020-01-09	54
10	서울	2020-01-10	73

```
> describe(seoul20_pm10[,3]) # 미세먼지농도 기술통계량량
    vars   n mean    sd median trimmed   mad min max range skew kurtosis   se
X1     1 299 46.1 23.13     42   43.79 19.27   8 120   112 1.01     0.96 1.34
```

❶ ❶ ❸





• 기술 통계량 분석:

❶ 평균과 중위 값 모두 46.1 ㎍/㎥, 43 ㎍/㎥ 로, 기준인 50 ㎍/㎥ 보다 낮다.

❷ 75%의 데이터가 60 ㎍/㎥보다 낮다.

❸ 최대 값은 120 ㎍/㎥ 로 50 ㎍/㎥ 보다 월등히 크며, 50 ㎍/㎥ 보다 큰 수치들도 다수 존재한다.

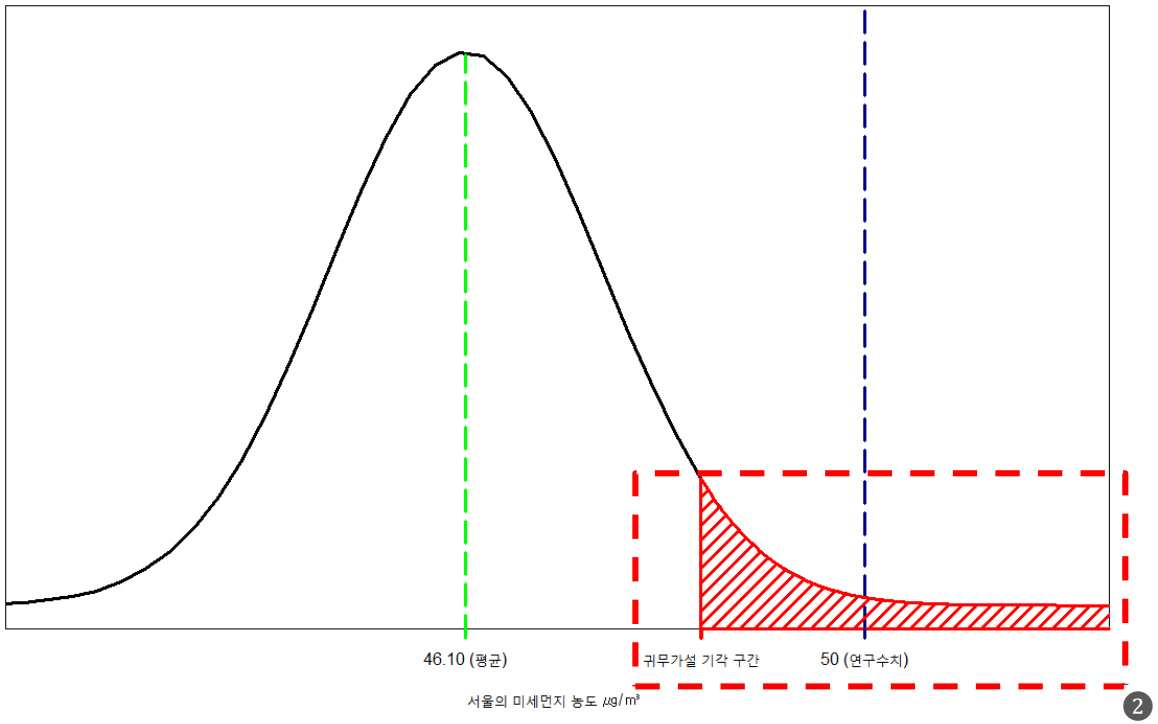
● 가설검정:

귀무가설 수립: 서울의 하루 평균 미세먼지 농도는 50 ㎍/㎥ 이상이다.
대립가설 수립: 서울의 하루 평균 미세먼지 농도는 50 ㎍/㎥ 미만이다.

```
        One Sample t-test

data:  seoul20_pm10$미세먼지농도
t = -2.9159, df = 298, p-value = 0.001908
alternative hypothesis: true mean is less than 50
95 percent confidence interval:
     -Inf 48.30702
sample estimates:
mean of x
 46.10033
```

❶

❷



● 통계량분석:

❶ P값이 유의수준인 0.05보다 낮음으로, 귀무가설을 기각할 통계적인 증거를 지닌다.
❷ 하루평균 46.1 ㎍/㎥에 대한 95% 신뢰구간은 48.31 ㎍/㎥ 이하이며, 초과 값들은 귀무가설을 기각하는 통계적인 증거를 지닌다.

● 최종결론:

2020년 12월 서울에서 측정한 평균 미세먼지 농도는 46.1 ㎍/㎥ 이다. 이는 유의수준 0.05의 가설검정을 통해서, 세계 보건기구에서 제공하는 하루평균 가이드라인인 50 ㎍/㎥ 보다 낮다고 할 수 있다.

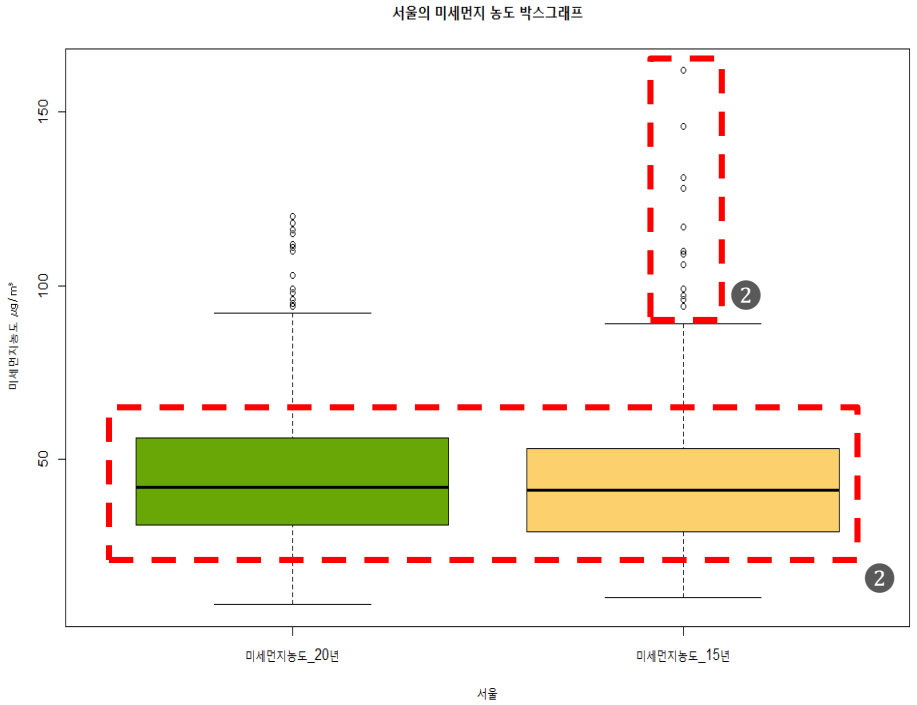
● 문제정의:

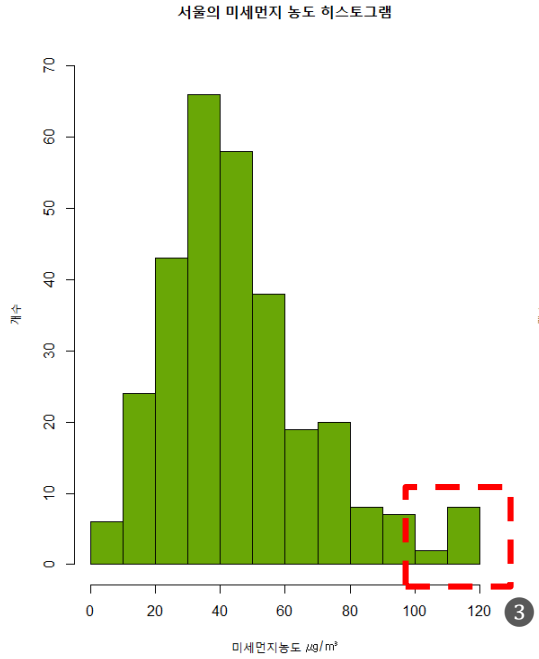
정부는 미세먼지를 줄이기 위해 2016년부터 1년의 단기대책과 2년의 중장기 대책을 세워 실행해 왔다.
그렇다면, 단기/중장기 대책이 실행되기 전인 2015년의 미세먼지 농도와, 대책이 실행되었던 2020년의 미세먼지 농도를 비교하여,
정부의 대책이 미세먼지를 줄이는 데 도움이 되었는지 알아보려 한다.

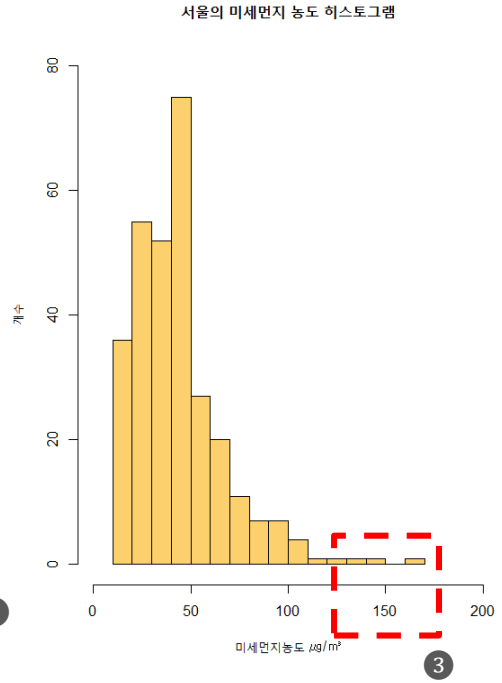
● 데이터 수집:

2016년/2020년 서울 매 일별 평균 미세먼지 농도 측정 데이터 / 출처: 통계청

```
> describe(df_pm10) # 20년/15년 기초통계량 분석
                vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
미세먼지농도_20년    1 299 46.10 23.13     42   43.79 19.27   8 120   112 1.01     0.96 1.34
미세먼지농도_15년    2 299 44.34 23.82     41   41.55 17.79  10 162   152 1.50     3.48 1.38
```







● 기술 통계량 분석:

1. 2020년과 2015년의 미세먼지 농도의 평균과 중위 값은 큰 차이를 보이지 않는다.
2. 박스그래프로부터 2020년과 2016년의 미세먼지 농도는 전반적으로 비슷함을 알수있지만, 2016년에는 비교적 많은 이상치가 존재한다.
3. 최대 값에서는 2020년과 2015년 큰 차이를 보이며, 2015년에는 비교적 큰 값들이 다수 존재한다.

● 가설검정:

귀무가설 수립: 2015년과 2020년의 미세먼지 농도의 차이는 없다.
대립가설 수립: 2015년과 2020년의 미세먼지 농도에 차이가 있다.

```
        Paired t-test

data:  df_pm10$미세먼지농도_20년 and df_pm10$미세먼지농도_15년
t = 0.95441, df = 298, p-value = 0.3407
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.868205  5.386599
sample estimates:
mean of the differences
               1.759197
```

❶

❷

● 통계량분석:

❶ P값이 유의수준인 0.05보다 큼으로, 귀무가설을 기각할 근거가 충분하지 않다.
❷ 95% 신뢰구간이 -1.87 에서 5.38이기 때문에, 2016년과 2020년의 평균차이인 1.76은 귀무가설을 유지한다.

● 최종결론:

앞서 기술통계 및 그래프 분석을 통해서 2015년과 2020년에는 어느정도 차이가 있음이 보여졌으나, 차이 검정을 통해 통계적으로는 그 차이가 유의미하다 판단하기 힘들다. 따라서, 정부의 미세먼지 단/중장기 대책이 미세먼지 농도를 줄이는데 통계적으로 성공했다고 말하기 힘들다.

• 문제정의:

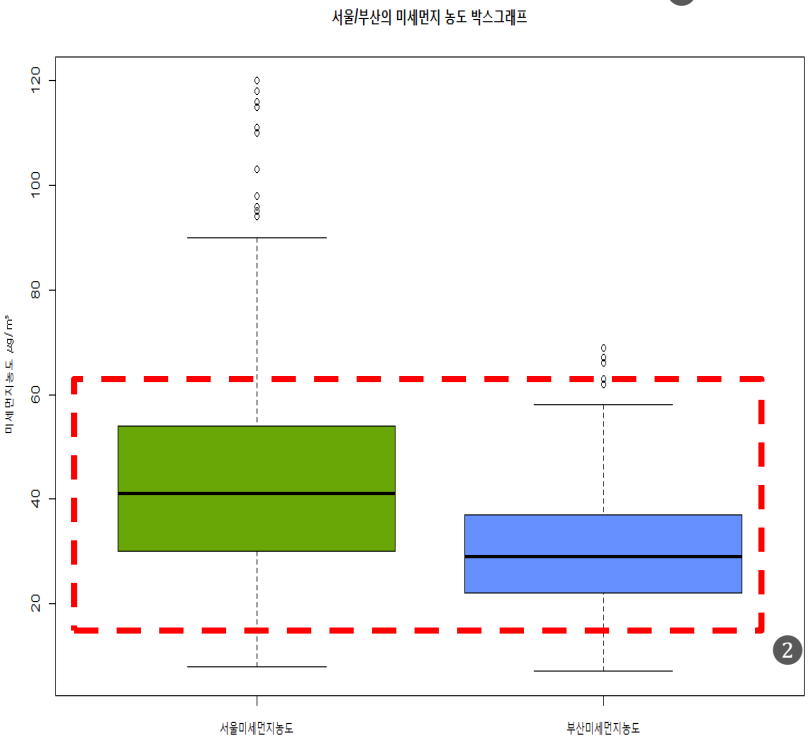
서울에서 멀리 떨어진 부산과 서울의 미세먼지 농도의 차이가 있는지 검정.

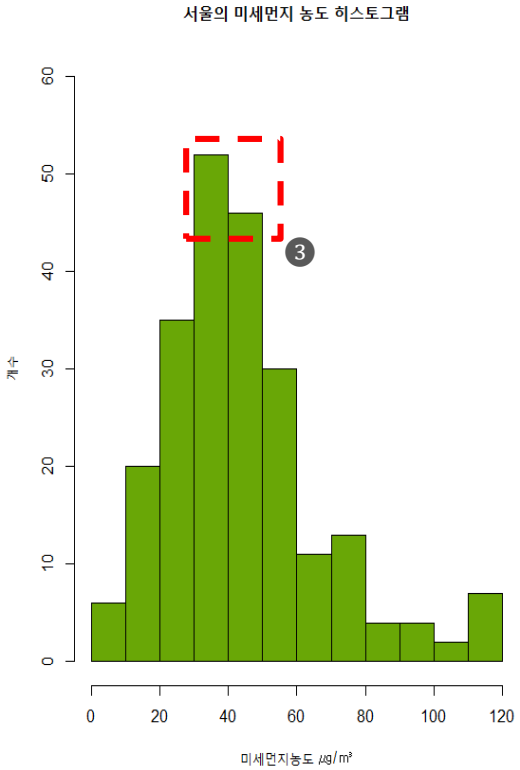
• 데이터 수집:

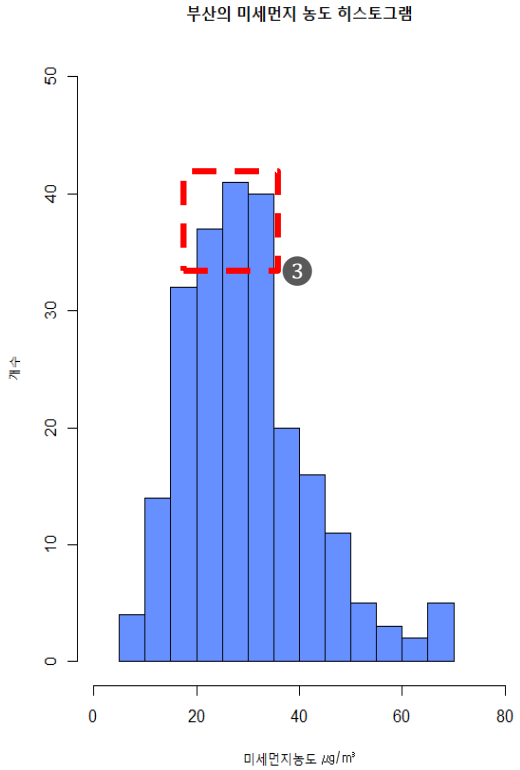
2020년 서울과 부산 매 일별 평균 미세먼지 농도 측정 데이터 / 출처: 기상청

```
> describe(df1_pm10)
                vars   n  mean    sd median trimmed   mad min max range skew kurtosis   se
서울미세먼지농도    1 230 44.81 23.14     41   42.22 17.79   8 120   112 1.16     1.49 1.53
부산미세먼지농도    2 230 30.47 12.24     29   29.41 10.38   7  69    62 0.86     0.75 0.81
```

① ①







• 기술 통계량 분석:

① 부산의 평균과 중위 값 모두 서울보다 낮다.

② 박스그래프에서 부산의 미세먼지 농도는 40 ㎍/㎥ 보다 낮게 분포된 반면, 서울은 30~55 ㎍/㎥ 로 비교적 넓게 분포 되어있다.

③ 히스토그램에서 서울의 미세먼지 농도는 30~50 ㎍/㎥ 가 가장 많았고, 제주의 경우 20~40 ㎍/㎥ 에 많이 분포 되어있다.

- 전체적으로 제주의 미세먼지 농도가 서울보다 낮게 보인다 .

● 가설검정:

귀무가설 수립: 서울과 제주의 미세먼지 농도는 같다.
대립가설 수립: 서울과 제주의 미세먼지 농도에는 차이가 있다.

```
> var.test(미세먼지농도 ~ 지역, data = df2_pm10)

        F test to compare two variances

data:  미세먼지농도 by 지역
F = 0.27972, num df = 229, denom df = 229, p-value < 2.2e-16
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2157481 0.3626636
sample estimates:
ratio of variances
         0.2797213
```

❶

```
        Welch Two Sample t-test

data:  미세먼지농도 by 지역
t = -8.3092, df = 347.82, p-value = 2.184e-15
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -17.73860 -10.94835
sample estimates:
mean in group 부산 mean in group 서울
         30.46957           44.81304
```

❷

● 통계량분석:

❶ 등분산검정 실행, P값이 0.05 보다 작음으로 등분산검정의 귀무가설을 기각하며, 이분산임을 확인
❷ 가설검정을 실행, P값이 0.05 보다 작음으로 귀무가설을 기각할 만한 통계적인 근거가 있다.

● 최종결론:

2020년 서울과 부산의 하루 평균 미세먼지 농도 수치는 부산이 서울보다 낮다는 것이 0.05 유의수준의 차이검정에 의해 입증되었다.

• 문제정의:

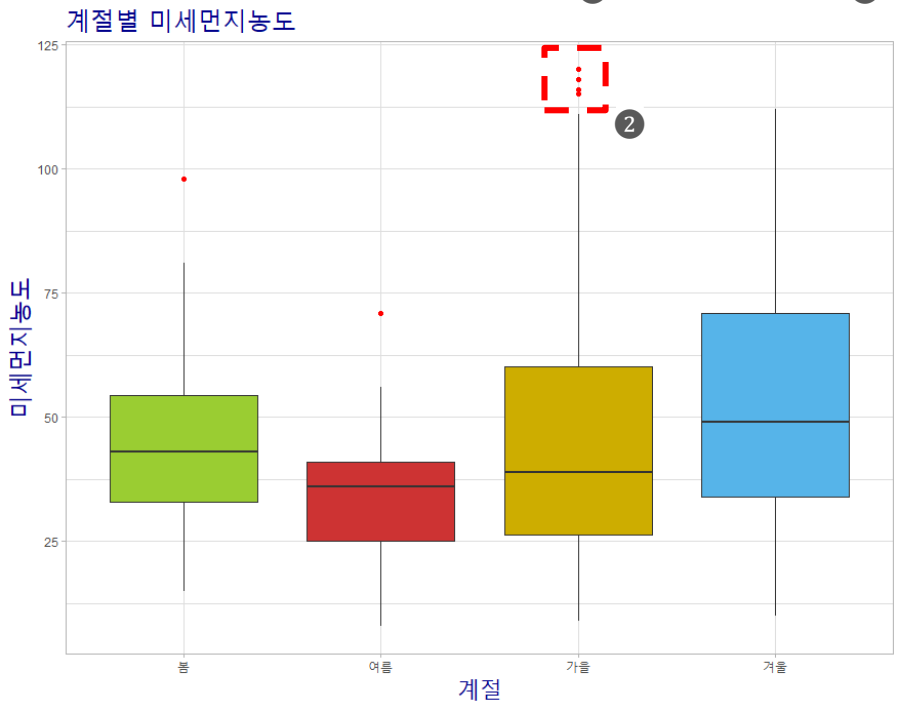
서울에는 사계절이 존재한다. 사계절을 각각 봄(3~5월), 여름(6~8월), 가을(9~11월), 겨울(12~2월)로 정의한다면, 계절별로 미세먼지 농도에 차이가 있는지 알아보려 한다.

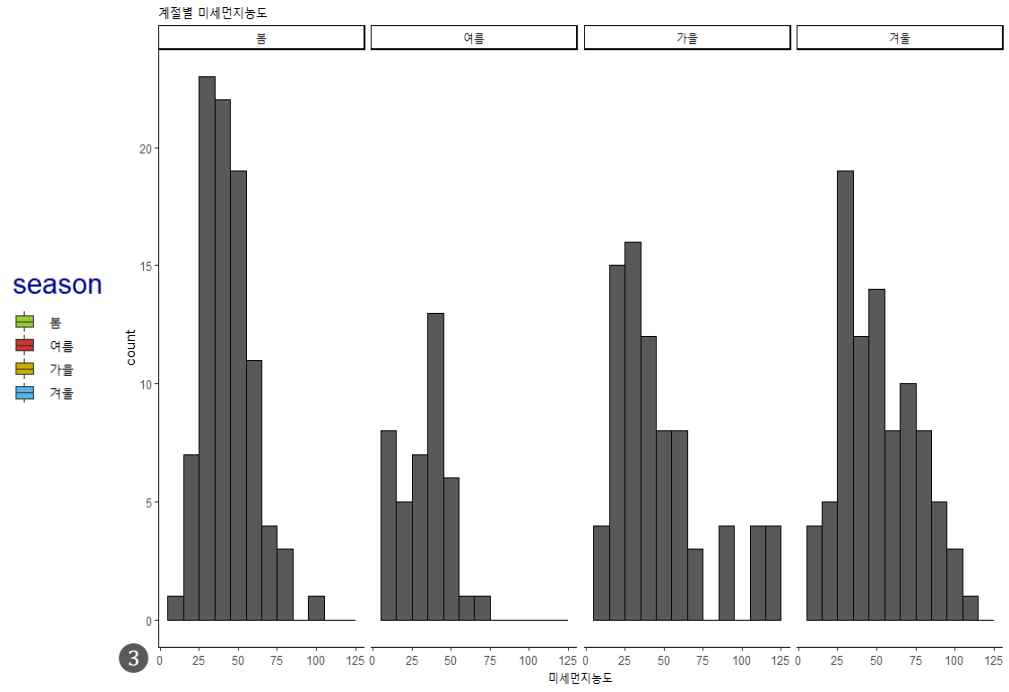
• 데이터 수집:

2020년 서울과 일별 평균 미세먼지 농도 측정 데이터 / 출처: 기상청

```
> describeBy(seoul_pm20_season$미세먼지농도, seoul_pm20_season$season, mat=T)
    item group1 vars  n     mean       sd median  trimmed     mad min max range       skew   kurtosis       se
X11    1     봄    1 91 44.40659 15.34563     43 43.47945 14.8260  15  98    83 0.68231728  0.6865554 1.608659
X12    2   여름    1 41 33.24390 14.39059     36 33.21212 14.8260   8  71    63 0.02201787 -0.3952299 2.247432
X13    3   가을    1 78 47.88462 30.00865     39 44.14062 22.2390   9 120   111 1.08231982  0.1858902 3.397810
X14    4   겨울    1 89 52.19101 23.82840     49 51.24658 26.6868  10 112   102 0.35937071 -0.6797139 2.525805
```

① ①





• 기술 통계량 분석:

① 여름의 평균 미세먼지 농도가 다른 계절에 비해 낮다. 반면 겨울의 평균 미세먼지 농도는 다른 계절에 비해 높다.

② 다른 계절에 비해, 가을에 많은 이상치가 발견 되었다.

③ 전체적으로 여름은 다른 계절에 비해 미세먼지 농도가 낮은 것으로 추정된다.

• 가설검정:

귀무가설 수립: 서울의 계절에 따른 미세먼지 농도에 차이는 없다.
대립가설 수립: 적어도 하나의 계절의 미세먼지 농도에는 차이가 있다.

```
#-------------------------------------------------------------------------
> bartlett.test(미세먼지농도 ~ season, data=seoul_pm20_season)

        Bartlett test of homogeneity of variances

data:  미세먼지농도 by season
Bartlett's K-squared = 48.777, df = 3, p-value = 1.455e-10

> #이분산일때 Welch's ANOVA test
> oneway.test(미세먼지농도 ~ season, data=seoul_pm20_season, var.equal = FALSE)

        One-way analysis of means (not assuming equal variances)

data:  미세먼지농도 and season
F = 11.439, num df = 3.00, denom df = 140.81, p-value = 9.283e-07
```

❶ ❷

```
#----Data Info-------------------------------
  Sample Size    Effect     Lower     Upper
1      봄   91 0.5354247 0.4947832 0.5756011
2    여름   41 0.3646241 0.3169135 0.4151521
3    가을   78 0.4994211 0.4484596 0.5503946
4    겨울   89 0.6005301 0.5551104 0.6442842
```

```
#----Analysis-------------------------------------------
      Estimator  Lower  Upper Statistic     p.Value
2 - 1    -0.171 -0.283 -0.054    -3.804 1.562327e-03
3 - 1    -0.036 -0.154  0.083    -0.794 8.558233e-01
4 - 1     0.065 -0.041  0.170     1.602 3.815175e-01
3 - 2     0.135 -0.003  0.267     2.571 5.650468e-02
4 - 2     0.236  0.110  0.354     4.866 2.488756e-05
4 - 3     0.101 -0.027  0.226     2.075 1.689573e-01
```

❸

• 통계량분석:

❶ 등분산검정 실행, P값이 0.05 보다 작음으로 등분산검정의 귀무가설을 기각하며, 이분산임을 확인
❷ 가설검정을 실행, P값이 0.05 보다 작음으로 귀무가설을 기각할 만한 통계적인 근거가 있다.
❸ 사후검정을 통해 모든 계절이 다른 계절과 차이가 있음에 대한 통계적인 근거가 있다.

• 최종결론:

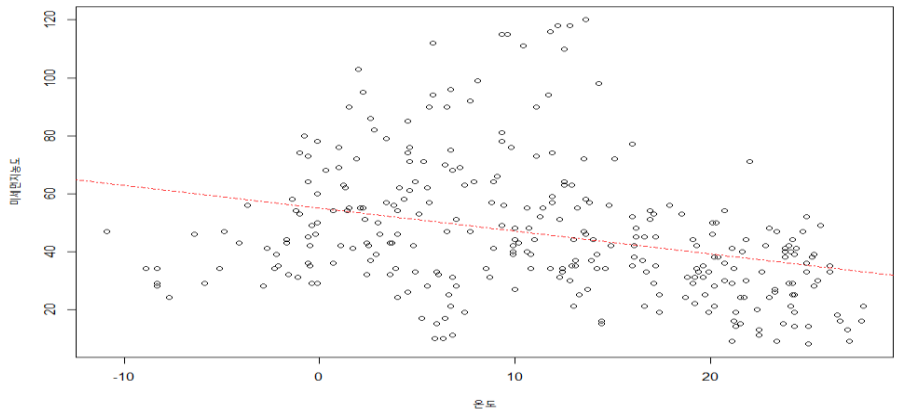
서울에서 계절에 따른 미세먼지 농도의 변화가 있음을 차이검정을 통해 통계적으로 유의미함이 확인되었다. 사후검정에서 보여지듯, 모든 계절은 다른 계절과 차이가 있다.

● 문제정의:

앞서 계절에 따라 미세먼지 농도에 어느정도 차이가 있음을 보았다. 그렇다면, 하루 평균온도와 그 날 평균 미세먼지 농도사이에 상관관계가 있는지 알아보려 한다.

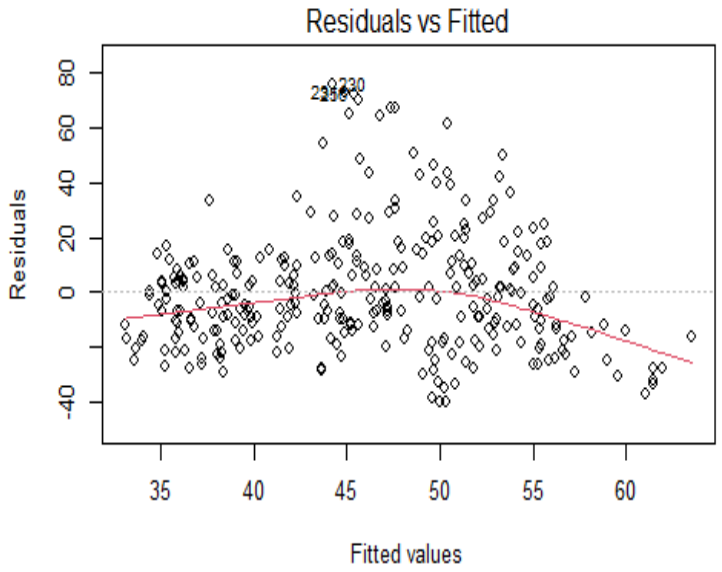
```
> summary(regModel)

Call:
lm(formula = 미세먼지농도 ~ 온도, data = seoul_pm20)

Residuals:
    Min      1Q  Median      3Q     Max
-40.313 -15.089  -3.089  10.650  75.741

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  54.9527     2.0282  27.094  < 2e-16 ***
온도          -0.7863     0.1402  -5.609 4.66e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.03 on 297 degrees of freedom
Multiple R-squared:  0.09578,   Adjusted R-squared:  0.09273
F-statistic: 31.46 on 1 and 297 DF,  p-value: 4.664e-08
```

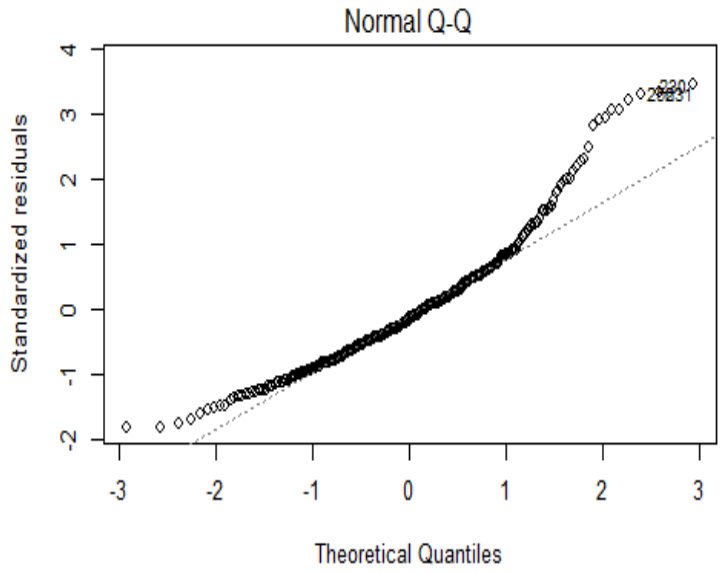
1 2 3 4 5

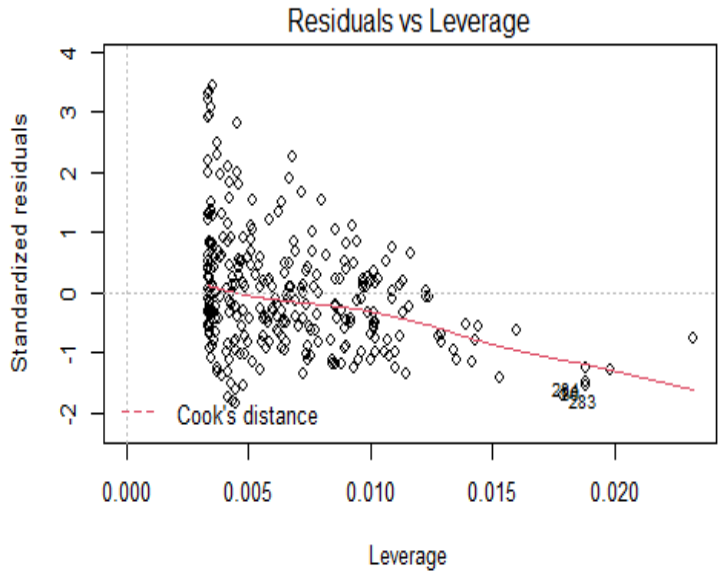


● 단순회귀분석:

1. 회귀모델 : 미세먼지농도(Y) = -0.7863온도(X) +54.9527
2. 설명변수인 온도의 계수 -.0.7863에 대한 P값이 0.05보다 작음으로 통계적으로 유의하다고 판단된다.
3. 온도가 1도 상승함에 따라, 미세먼지 농도는 0.7863 만큼 감소한다.
4. 온도는 미세먼지 농도의 차이 변화에 대략 9.7%의 설명력을 갖는다.
5. P값이 0.05 보다 작음으로, 현재 온도와 미세먼지 농도 사이의 단순 회귀 모델식은 통계적으로 유의하다고 판단되나, 온도의 변화는 미세먼지 농도 변화의 9.7% 정도에만 유의하다 판단됨으로, 회귀모델의 변형 혹은 다른 설명변수가 추가적으로 필요하다 판단된다.







- **<u>단순회귀분석 가정진단:</u>**

- 잔차들이 고르게 분포되어 있음으로, 등분산의 가정을 만족한다. 그러나 잔차들의 패턴이 곡선을 형성함으로 추후 설명변수를 2차 함수 등으로 시도 가능성 있음.

- +2를 넘어가는 이상치 등이 다수 존재하며 정규분포선에서 많이 벗어나 있음으로 정규성을 완벽하게 따른다 말하기 힘들다.

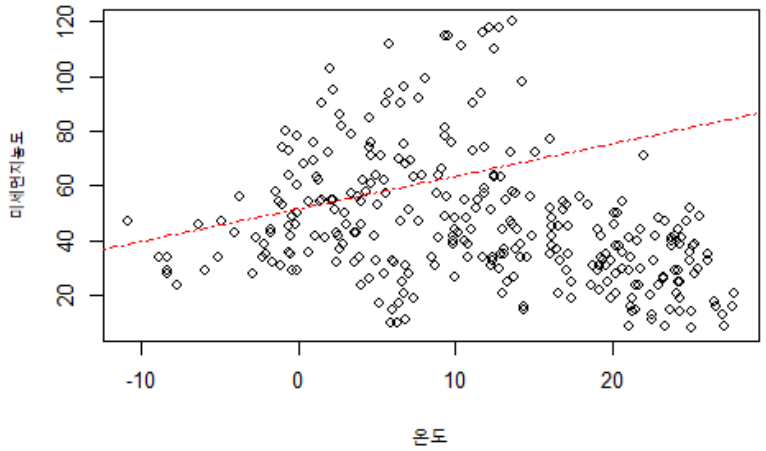
- 관측 값 번호 230, 283번 등의 이상치가 존재.

● 문제정의:

앞서 단순회귀분석을 설명력을 보안하는 새로운 모델식을 세우기 위해 다항회귀분석 시도.

```
> summary(regModel)

Call:
lm(formula = 미세먼지농도 ~ 온도 + I(온도^2), data = seoul_pm20)

Residuals:
    Min      1Q  Median      3Q     Max
-45.497 -13.687  -2.246  10.818  68.915

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.57505    1.97936  26.056  < 2e-16 ***
온도         1.19074    0.33851   3.518 0.000504 ***
I(온도^2)   -0.09020    0.01423  -6.340 8.54e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.7 on 296 degrees of freedom
Multiple R-squared:  0.2039,    Adjusted R-squared:  0.1985
F-statistic: 37.91 on 2 and 296 DF,  p-value: 2.204e-15
```

❶ ❷ ❸ ❹



● 다항회귀분석:

❶ 회귀모델 : 미세먼지농도(Y) = 1.19온도(X) -0.09온도^2(X^2) +51.57

❷ 설명변수인 온도와 온도^2의 계수의 P값 모두 0.05 모다 작음으로 통계적으로 유의미하다고 판단 할 수 있다.

❸ 온도는 미세먼지 농도의 차이 변화에 대략 20%의 설명력을 갖는다.

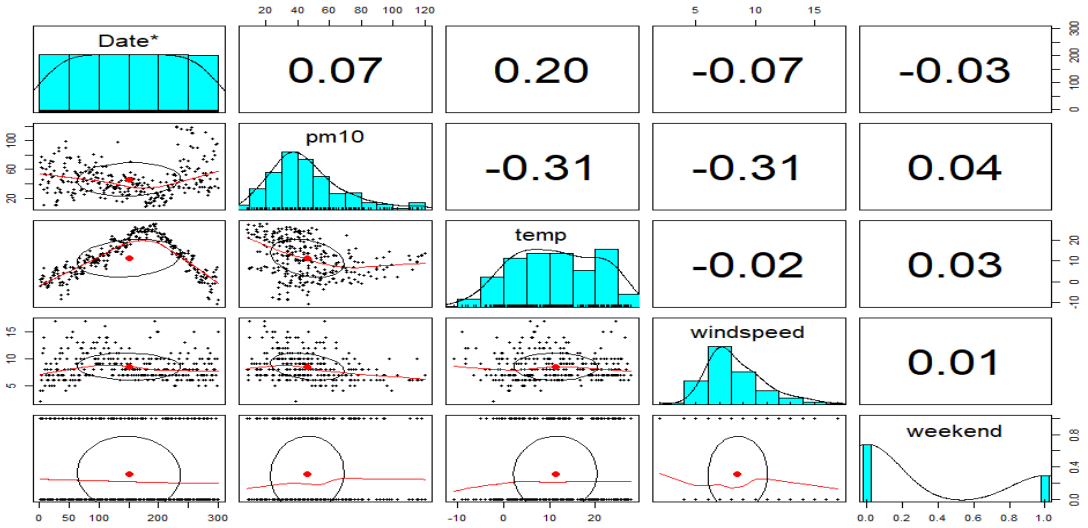
❹ P값이 0.05 보다 작음으로, 현재 온도와 미세먼지 농도 사이의 다항 회귀 모델식은 통계적으로 유의하다고 판단된다

● 최종결론:

- 다항회귀분석의 모델식은 통계적으로 유의미하다 판단되고 설명력 역시 단순회귀분석에 비해 증가 했지만, 아직 미세먼지 농도의 변화를 온도의 차이만으로 설명하기에는 현저히 부족하다고 판단된다.

● 문제정의:

앞서 다항회귀분석에 다른 설명변수들을 추가하여 미세먼지 농도에 영향을 주는 변수를 확인한다 .
미세먼지는 바람의 영향을 받음으로 풍속과, 공장 등 미세먼지를 유발하는 산업활동이 주말에 쉰다는 가정하에
추가 변수: 평균 풍속(m/s), 주말여부(0-주중, 1-주말/공휴일) 를 추가하였다.




```
Call:
lm(formula = pm10 ~ temp + I(temp^2) + windspeed + weekend, data = seoul_pm20)

Residuals:
    Min      1Q  Median      3Q     Max
-42.778 -13.335  -1.517  10.351  64.772

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 76.61831    4.28655  17.874  < 2e-16 ***
temp         1.21970    0.31674   3.851 0.000145 ***
I(temp^2)   -0.09237    0.01331  -6.940 2.51e-11 ***
windspeed   -2.98479    0.44551  -6.700 1.06e-10 ***
weekend      1.68390    2.41271   0.698 0.485775
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.34 on 294 degrees of freedom
Multiple R-squared:  0.3102,    Adjusted R-squared:  0.3008
F-statistic: 33.05 on 4 and 294 DF,  p-value: < 2.2e-16
```

❶

❷

● 다중회귀분석:

❶ 풍속의 P값은 0.05보다 작지만 주말여부의 P값은 0.05보다 큼으로, 풍속은 미세먼지 농도에 영향을 주는 설명변수로써 통계적으로 의미가 있지만, 주말여부는 그렇지 못하다.

❷ 앞선 단순/다항 회귀분석에 비해 10%정도 증가한 대략 30%의 설명력을 지닌다.

● 최종결론:

- 미세먼지 농도는 온도와 풍속에 영향을 받아 온도, 풍속이 높아지면 미세먼지 농도가 증가함을 보였다. 반면, 주말/공휴일을 나눈 요일의 여부는 미세먼지 농도에 통계적으로 영향을 끼쳤다고 보기 힘들다.