<데이터 분석 응용>

미세먼지 농도와 중국 코로나-19

-2개년 마포구 미세먼지 농도 비교와 중국 코로나 확진자 수 중심으로

학과: 산업공학과

학번: B6XXXXX

이름: 김준석

<목차>

<그림목차>

1. 서론

2020년 1월에 시작한 코로나-19가 우리 사회에 많은 변화를 일으켰다.

길거리에 나가면 10명중 9명은 마스크를 착용하고 있고, 대부분의 학교들은 원격으로 강의를 실시하고 있다.

많은 변화들 중에서 이로운 변화는 대한민국 하늘에 미세먼지가 최근 2~3년 대비 적다는 점이다. 2019년에는 한 겨울임에도 불구하고, 미세먼지가 심각해 하늘이 누렇게 보이는 현상이 있었으며, 특히 3월에는 황사와 함께 유입되어 숨쉬기 매우 힘든 상황도 발생하였다. 하지만 2020년 미세먼지가 심각 수준까지 되었던 적이 손에 꼽힐 만큼 매우 깨끗한 하늘을 유지하고 있다.

코로나-19가 중국에서 터진 때마침, 미세먼지가 잠잠해져 각 종 SNS나 포털 사이트에서 코로나-19로 인하여 중국 공장의 가동률이 줄어 미세먼지 농도가 줄었다는 말이 나왔다. 물론 중국의 막대한 공장 가동으로 인하여 미세먼지가 대한민국에 영향을 주는 것을 알았지만, 이 문제가 단지 중국 공장의 가동률로만 해결될 수 있는 문제인지 궁금하였다.

그리하여 이번 프로젝트 주제로 중국 코로나-19 발병수와 대한민국 마포구 미세먼지 농도 간의 인과 관계가 있는 지 파악하고자 한다.

2. 분석

2-1. 사용한 데이터

이 주제를 분석하기 위해 2019년 미세먼지 수치, 2020년 미세먼지 수치와 중국코로나-19 확진자 수의 데이터가 필요했다.

2019년과 2020년 미세먼지 수치는 각 1월 1일부터 4월 17일로 하였다. 그 이유는 중국에서 시작된 코로나-19는 2019년 12월말부터 각 종 뉴스에서 보도가 되었기 때문에 각 해 1월 1일을 기준으로 하였으며, 4월 17일이후로는 확진자 수가 30명미만이여서 4월 17일까지 범위를 정하였다. 중국 코로나-19 확진자 수 데이터는 1월 21일부터 4월 17일로 정하였는데, 중국에서 공식 인정한 날이 1월 21일이라 그 이전의 데이터가 없어 1월 21일부터 4월 17일까지 확진자 수를 가지고 분석한다.

2019년과 2020년 미세먼지 데이터는 '서울 열린 데이터 광장' 에서, 중국 코로나-19 확진자 데이터는 'coronaboard' 라는 사이트에서 얻었다.

2-2. 분석도구 및 분석 절차

프로젝트는 2019년도와 2020년도 미세먼지 수치의 차이가 있는가를 확인하고, 차이가 있으면 2020년도 미세먼지 수치와 중국 코로나-19 확진자 수간의 회귀분석을 실시한다

데이터 조작은 'dplyr' 패키지를 사용한다. 2019년 미세먼지 농도와 2020년 미세먼지 농도를 그래프로 비교하기 위해 'ggplot2' 패키지를 사용했다.

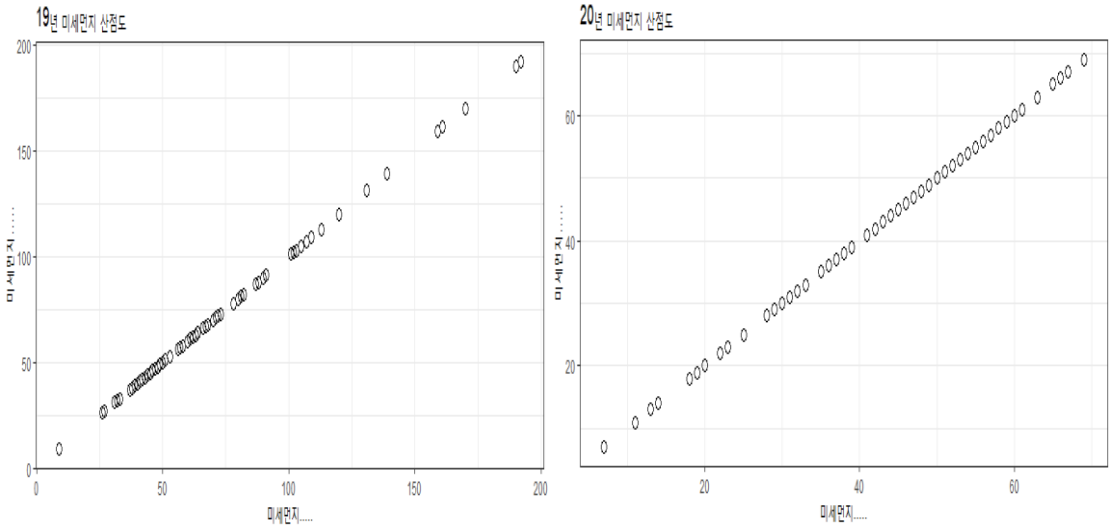
1) 데이터 불러오기

```
> library(dplyr)
> library(ggplot2)
> library(ggExtra)
>
>
> #csv불러오기
> finedust_19=read.csv('C:\\Users\\ie gram_08\\Desktop\\19년미세먼지.csv',stringsAsFactor = FALSE,encoding="utf-8")
> finedust_20=read.csv('C:\\Users\\ie gram_08\\Desktop\\20년미세먼지.csv',stringsAsFactor = FALSE,encoding="utf-8")
> str(finedust_19)
'data.frame':   100 obs. of  8 variables:
 $ 측정일자           : chr  "1월01일" "1월02일" "1월03일" "1월04일" ...
 $ 측정소명           : chr  "마포구" "마포구" "마포구" "마포구" ...
 $ 미세먼지.....      : int  42 38 38 67 70 57 62 47 56 58 ...
 $ 오존.ppm.          : int  18 18 19 41 39 22 37 20 21 31 ...
 $ 이산화질소농도.ppm.: num  0.014 0.009 0.007 0.003 0.013 0.01 0.007 0.016 0.007 0.005 ...
 $ 일산화탄소농도.ppm.: num  0.022 0.029 0.029 0.047 0.024 0.026 0.037 0.021 0.027 0.032 ...
 $ 아황산가스농도.ppm.: num  0.5 0.6 0.6 1.1 0.6 0.5 0.8 0.5 0.6 0.6 ...
 $ 초미세먼지.....    : num  0.003 0.003 0.002 0.004 0.003 0.004 0.004 0.004 0.004 0.004 ...
> summary(finedust_19)
   측정일자           측정소명          미세먼지.....      오존.ppm.       이산화질소농도.ppm.  일산화탄소농도.ppm.  아황산가스농도.ppm.  초미세먼지.....
 Length:100         Length:100         Min.   :  9.00   Min.   :  6.00   Min.   :0.00000   Min.   :0.00000   Min.   :0.000   Min.   :0.00000
 Class :character   Class :character   1st Qu.: 43.00   1st Qu.: 20.75   1st Qu.:0.01275   1st Qu.:0.02100   1st Qu.:0.400   1st Qu.:0.00300
 Mode  :character   Mode  :character   Median : 59.00   Median : 29.00   Median :0.02050   Median :0.02750   Median :0.500   Median :0.00400
                                       Mean   : 67.54   Mean   : 37.74   Mean   :0.01987   Mean   :0.03082   Mean   :0.597   Mean   :0.00398
                                       3rd Qu.: 80.25   3rd Qu.: 44.25   3rd Qu.:0.02625   3rd Qu.:0.03925   3rd Qu.:0.700   3rd Qu.:0.00425
                                       Max.   :192.00   Max.   :141.00   Max.   :0.04400   Max.   :0.05900   Max.   :1.400   Max.   :0.00600
> summary(finedust_20)
   측정일자           측정소명          미세먼지.....    초미세먼지.....    오존.ppm.       이산화질소농도.ppm.  일산화탄소농도.ppm.  아황산가스농도.ppm.
 Length:105         Length:105         Min.   : 7.00   Min.   : 4.00   Min.   :0.00300   Min.   :0.01200   Min.   :0.1000   Min.   :0.002000
 Class :character   Class :character   1st Qu.:31.00   1st Qu.:17.00   1st Qu.:0.01300   1st Qu.:0.02300   1st Qu.:0.4000   1st Qu.:0.003000
 Mode  :character   Mode  :character   Median :42.00   Median :27.00   Median :0.02000   Median :0.03300   Median :0.5000   Median :0.003000
                                       Mean   :41.02   Mean   :27.68   Mean   :0.01962   Mean   :0.03352   Mean   :0.5019   Mean   :0.003314
                                       3rd Qu.:51.00   3rd Qu.:36.00   3rd Qu.:0.02500   3rd Qu.:0.04200   3rd Qu.:0.6000   3rd Qu.:0.004000
                                       Max.   :69.00   Max.   :59.00   Max.   :0.04200   Max.   :0.06100   Max.   :1.2000   Max.   :0.005000
>
> #dplyr 설정
> finedust_19_df=tbl_df(finedust_19)
> finedust_20_df=tbl_df(finedust_20)
```

<그림2-1> 데이터 불러오기

그림 <2-1>과 같이 csv파일의 데이터를 R에 불러와 'dplyr' 패키지를 이용해 데이터 프레임을 설정한다.

2) 산점도 그리기



<그림2-2> 2019년도, 2020년도 미세먼지 산점도

먼저 개략적으로 데이터를 파악하기 위해 2019년도 미세먼지 농도에 대한 산점도와 2020년도 미세먼지 농도에 대한 산점도를 그린다. <그림2-2>와 같이 산점도가 겉보기에는 차이가 없어 보이나 각 축을 보면 2019년도는 축의 범위가 0부터 200까지, 2020년도는 0부터 67까지 설정됐다. 이를 통해 2019년보다 2020년 미세먼지 농도가 개선됐다는 점을 알 수 있다.

3)차이검정

```
> #차이 검정
> x1=x[,1:2]
> y1=y[,1:2]
> xy=merge(x1,y1,by='측정일자',all=T)
> xy[!complete.cases(xy),]
   측정일자 미세먼지......x 미세먼지......y
53   2월22일                72                NA
54   2월23일                64                NA
68   3월09일                NA                60
69   3월10일                NA                44
70   3월11일                NA                32
71   3월12일                NA                32
72   3월13일                NA                29
75   3월16일                NA                23
76   3월17일                NA                45
> xy=na.omit(xy)
> xy2=xy[,2:3]
> chisq.test(xy2)

        Pearson's Chi-squared test

data:  xy2
X-squared = 1012.6, df = 97, p-value < 2.2e-16
```

<그림2-3> 차이검정

미세먼지 데이터가 일부 누락된 날도 있어, 2019년도 데이터와 2020년도 데이터를 '측정일자' 기준으로 합병하여 NA값을 제거한다. 그 후 카이제곱을 이용하여 차이 검정을 실시한다. P-value가 0.05보다 작으므로 귀무가설을 기각한다. 따라서 19년도 미세먼지 농도와 20년도 미세먼지 간의 차이가 존재한다고 볼 수 있다.

4)상관분석과 회귀분석

```
> cor(co_20$미세먼지.....,co_20$중국.코로나)
[1] 0.0726585
> model=lm(co_20$미세먼지.....~ co_20$중국.코로나)
> summary(model)

Call:
lm(formula = co_20$미세먼지..... ~ co_20$중국.코로나)

Residuals:
     Min       1Q   Median       3Q      Max
-29.1155  -9.0925  -0.7344   9.2487  29.1537

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       3.971e+01  1.702e+00  23.325   <2e-16 ***
co_20$중국.코로나 5.278e-04  7.953e-04   0.664    0.509
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.06 on 83 degrees of freedom
Multiple R-squared:  0.005279,  Adjusted R-squared:  -0.006705
F-statistic: 0.4405 on 1 and 83 DF,  p-value: 0.5087
```

<그림2-4> 상관분석과 회귀분석

중국 코로나-19 확진자 수 엑셀을 불러 2020년도 데이터와 합쳤다. 그 과정은 생략한다.

회귀분석 전 상관분석을 실시한다. 상관분석 실시한 결과 상관계수가 0.72로 강한 양의 상관관계를 나타내고 있다.

그 다음으로 선형회귀분석을 실시한다. 선형회귀분석을 한 결과 p값이 0.05보다 커 귀무가설을 채택한다. 따라서 미세먼지와 중국 코로나-확진자 수는 서로 인과관계가 없다고 할 수 있다.

5. 결론

이번 분석을 통하여 미세먼지 농도와 코로나-19와는 관계가 없다는 결론을 도출할 수 있다.

처음 주제를 선정했을 당시 중국을 비하하는 발언과 함께 '우리나라 미세먼지는 다 중국 때문이다.' 는 말이 많았고, 실제로 2020년 코로나가 중국에 강타하고 중국 공장들이 일시적으로 가동을 중단하자, 미세먼지 농도가 낮아졌다. 그리하여 이번 프로젝트의 결론이 당연하게도 '회귀관계가 있다.' 라는 결론이 도출이 될 줄 알았다.

하지만 분석 결과 예상을 뒤집는 결론이 도출됐다. 분석 이후 2020년 미세먼지 농도가 낮은 원인을 찾아본 결과 정부의 계절 관리제와 코로나-19 영향, 기상 조건이 더해진 것이라고 한다.

어느 정도 영향이 있지만 이번 분석함에 있어서 정부의 계절 관리제와 기상조건이 변수로 작용하여 아예 관계가 없는 것처럼 결론이 도출되었다. 이 변수는 연구자가 직접 통제할 수 없는 변수이다.

따라서 더 좋은 결과를 도출하기 위해서는 코로나-19가 지속되면서, 기상 조건이 작년과 균일하고, 계절 관리제가 잘 이행되지 않아야 한다. 이러한 조건이 작용되기 위해서는 긴 시간 동안의 수집 데이터가 필요하다. 또한 중국 코로나 확진자 수도 지속적으로 증가해야 한다.