**데이터구조와 컴퓨팅**

<오존에 따른 대기오염의 관계>



|  |  |
| --- | --- |
| **과목** | 데이터구조와 컴퓨팅 |
| **담당교수** | 손대순 교수님/김윤태 교수님 |
| **학과,이름** | 데이터테크전공 20183226신희진 |
| **제출일** | 2020.12.11 |

**목 차**

**Ⅰ. 서론**

1. 분석 개요
2. 분석 의미

* 분석 이유

**Ⅱ. 본론**

1. 데이터 설명
2. 분석 코드 정리

-기초통계량 분석

-상관계수

**Ⅲ. 결론**

1. 데이터분석 결과 및 해석

-산점도

1. 새로 알게 된 사실

**Ⅳ. 출처**

**I. 서 론**

**ⅰ. 분석 개요**

Airquality의 데이터를 사용하여, 변수 6개 중 Ozone(오존의 농도)와 Solar.R(일사량), Wind(풍속), Temp(온도) 4개 만 사용하여 기초통계량과 상관분석을 할 것이다. 각 평균과 분산, 표준 편차를 확인하여 변수 간의 연관된 정도를 살펴보고 해석을 한다.

**ⅱ. 분석 의미**

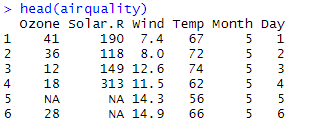
-분석 이유

어떤 데이터를 사용하여 상관분석을 해보지 고민 중, r프로그래밍에서 기본으로 제공해주는 데이터 airquality을 발견하였다. 실제 전광판에 기록된 오존과 풍속 등이 적혀있는 데이터였다. 오존의 농도가 일사량과의 연관성이 있는지, 오존의 농도가 풍속과 온도에 영향이 있는지가 궁금하여 airquality의 데이터를 선택하게 되었던 것 같다. 미세먼지가 많아지는 오늘 날, 대기환경지수에 대해 각 데이터들의 서로 간 연관성에 대해 알아볼 예정이다.

**Ⅱ. 본 론**

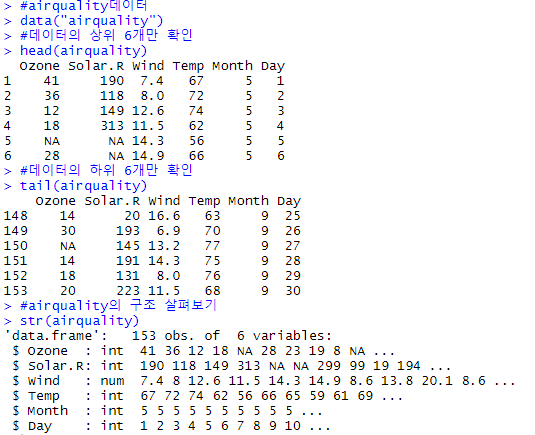
**ⅰ. 데이터 설명**

사용할 데이터는 뉴욕의 1973년도 공기의 질을 측정한 airquality는 R에서 기본으로 제공해주는 데이터로 대기환경지수의 세계 대기 오염지수를 담고 있다. 변수로는 Ozone(오존의 농도)와 Solar.R(일사량), Wind(풍속), Temp(온도), Month, Day 의 정보가 담겨있다.

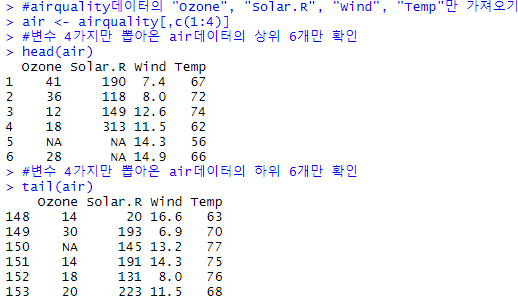


**ⅰⅰ. 분석 코드 정리**

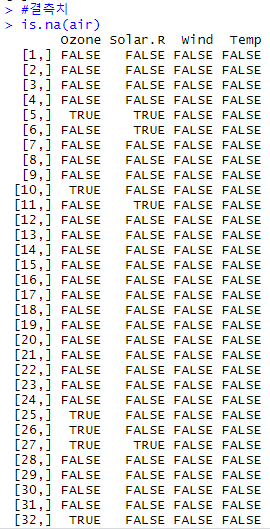
airquality를 사용하기 위해 어떤 데이터들이 들어있는지 먼저 확인을 한다. head함수와 tail함수를 사용하여 상위,하위 6개를 확인해보고 str을 사용하여 데이터의 구조를 확인해본다.



airquality의 데이터를 확인해본 결과, 153개의 데이터와 6가지의 변수를 가지고 있는 것을 알 수 있다. “Month”와 “Day”는 상관계수를 분석하기에는 적절하지 않다고 생각이 들어 2가지를 제외한 나머지의 변수만을 사용하기 위해 새로운 air에 데이터를 저장해주고, 다시 확인을 해준다.

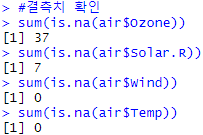


head를 사용하여 데이터를 확인해보았을 때, 결측치가 발견되는 것을 볼 수 있다. 결측치가 있는지 is.na를 사용하여 다시 확인을 해보았다.



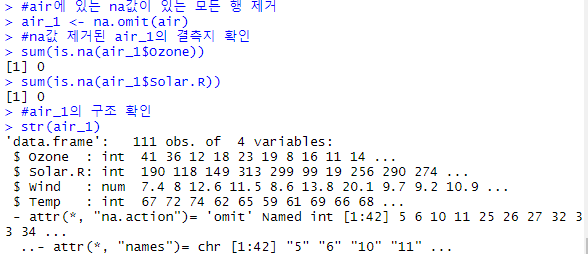
* TRUE로 나오는 값이 na값이라는 것을 알 수 있다.

각각의 변수에 결측치가 몇 개씩 있는지 확인을 하기 위해 sum함수를 사용하여 na값을 확인한다.



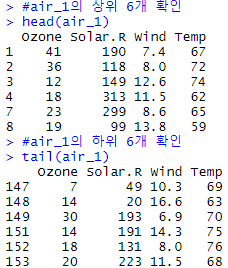
* 확인결과 “Ozone”과 “Solar.R”에서만 결측치가 발견된 것을 확인할 수 있었다.

air에 있는 데이터들에서 결측치가 발견된 두 개의 데이터들을 na.omit을 사용하여 na값이 포함된 모든 행들을 제거시켜준다. 그 후 결측치가 있는지 다시 확인을 해주고, 구조를 확인한다.



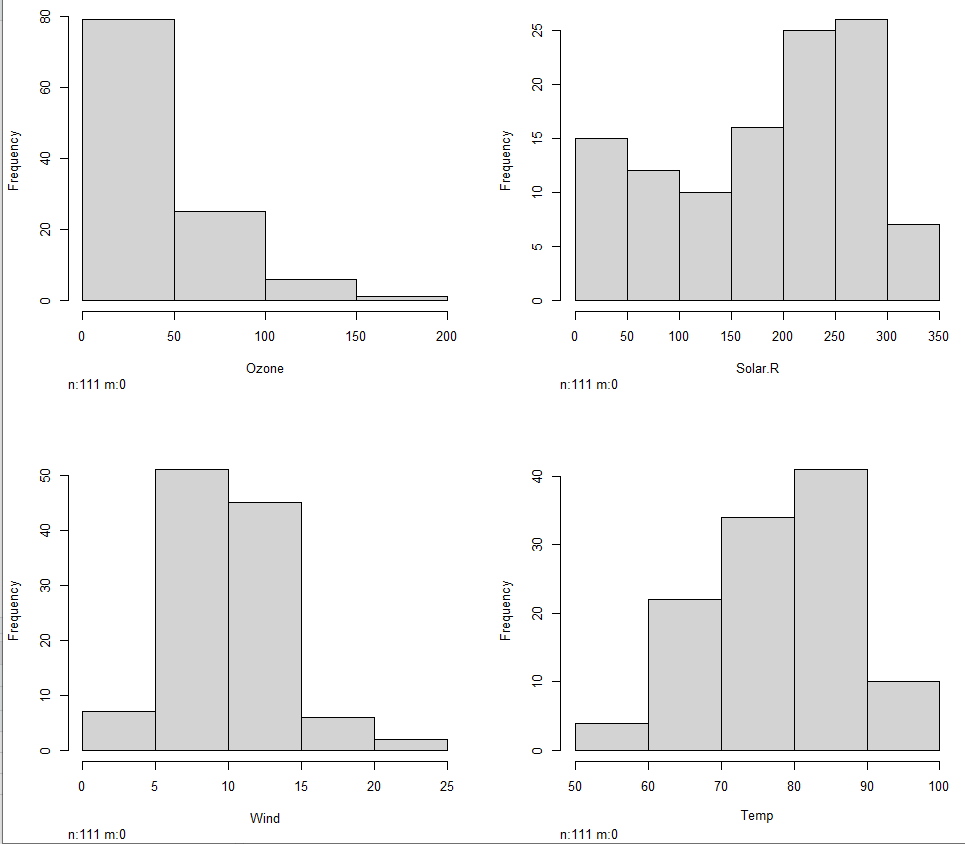
* 결측치를 제거해주고 데이터를 다시 확인해본 결과, 결측치가 완전히 제거되었다는 사실을 알 수 있다. 또한 결측치가 있는 모든 행을 지워서 저장해준 air\_1의 데이터의 구조를 살펴보면 4개의 변수에 111개의 데이터들이 저장되어있는 사실을 알 수 있었다.

na값을 제거해준 air\_1데이터의 상위/하위 6개씩 확인해보았다.



결측치를 제거해준 ari\_1을 사용하여 히스토그램을 그려보았다. 히스토그램은 연속형 자료에 대한 도수분포표를 시각화한 그래프다.

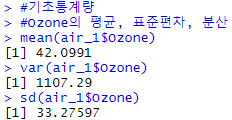




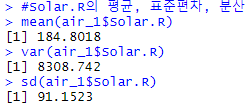
* x축은 각각의 변수들을 나타내고, y축은 빈도를 나타내었다. “Ozone”에서는 0~50사이의 값이 가장 많이 나타나며 다른 값들에 비해 확연하게 많은 것을 알 수 있었다. “Solar.R”에서는 250~300의 값이 가장 많았지만, 200~250의 값과 근소한 차이를 나타내었다. “Wind”에서는 5~10의 값이 가장 많았고, 10~15의 값과 근소한 차이를 나타내었지만 그 외의 값들과 많은 차이를 보였다. “Temp”에서는 80~90의 값이 가장 많은 것을 볼 수 있다.

-기초통계량 분석

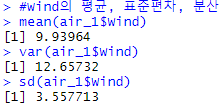
기초통계량이란 전체 데이터의 특성을 표현하는 수치를 뜻한다. Airquality의 데이터에서는 평균, 표준편차, 분산, 최솟값 등을 살펴 보려고 한다.



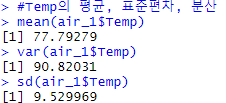
* na값을 제거한 “Ozone”의 기초통계량을 나타낸 것이다.



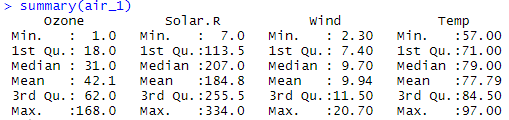
* na값을 제거한 “Solar.R”의 기초통계량을 나타낸 것이다.



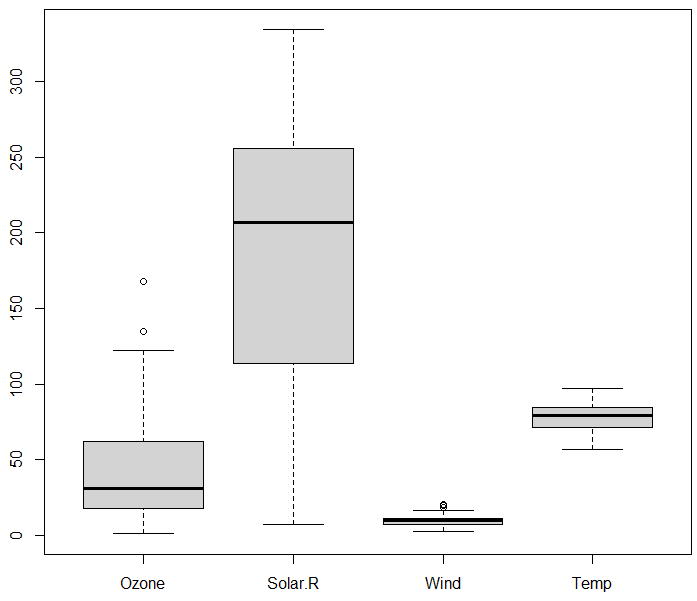
* na값을 제거한 “Wind”의 기초통계량을 나타낸 것이다.



* na값을 제거한 “Temp”의 기초통계량을 나타낸 것이다.

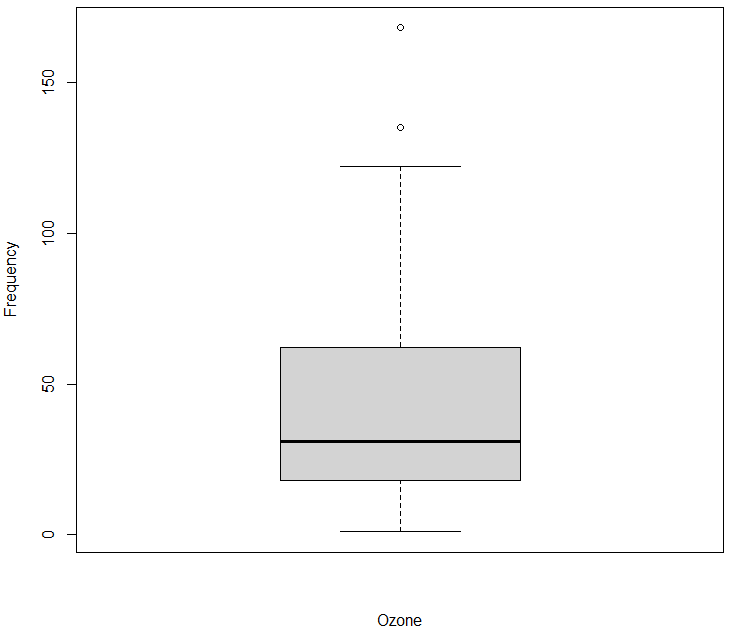


* summauy함수를 사용해서 4개의 변수를 한번에 보여주었다. summary에는 최소값, 제 1사분위수, 중앙값, 평균 제 3사분위수, 최대값을 보여준다.



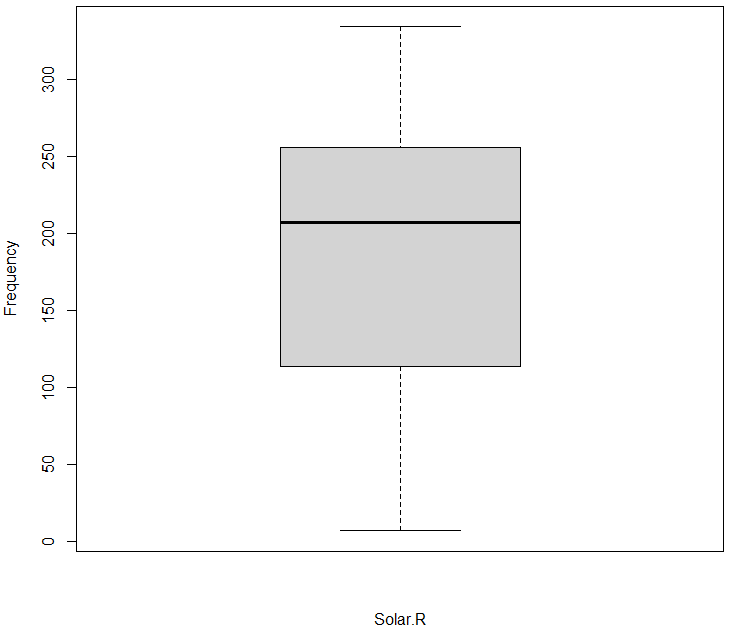
위의 데이터들을 가지고 상자그림을 그려보았다. 한번에 4개의 변수의 상자그림을 그렸지만 수치차이가 너무 많이 나서 그래프를 자세히 보기 어려웠다. 그래서 각각의 변수 하나하나의 그래프를 따로 그려보았다.





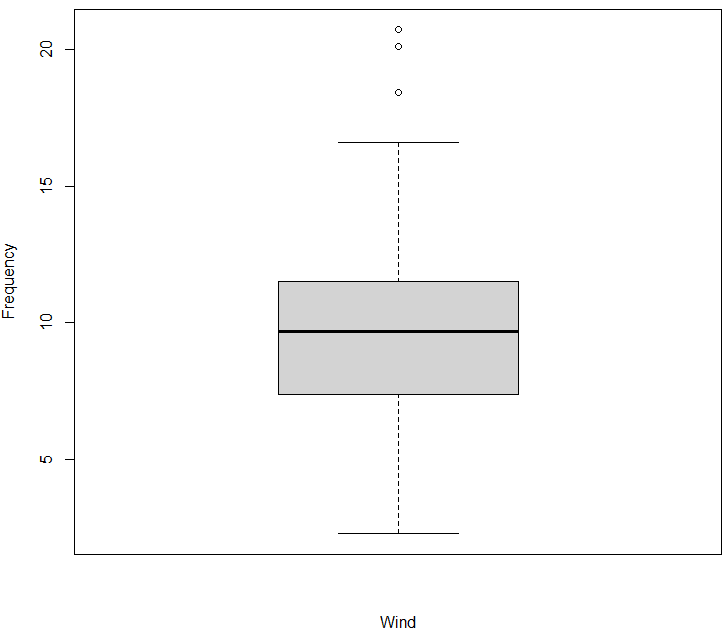
* “Ozone”에 대한 상자그림이다. 기초통계량을 구한 것과 같은 값의 데이터로 그려진 것을 확인할 수 있다.





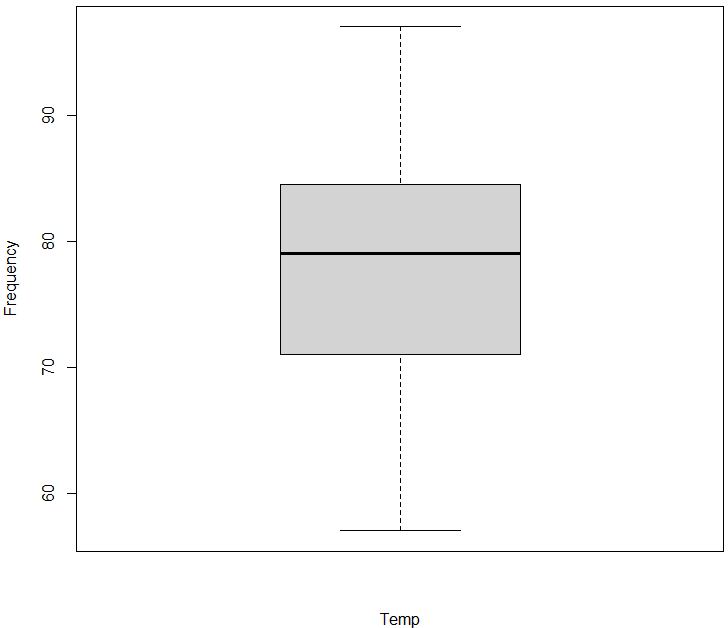
* “Solar.R”에 대한 상자그림이다. 기초통계량을 구한 것과 같은 값의 데이터로 그려진 것을 확인할 수 있다.





* “Wind”에 대한 상자그림이다. 기초통계량을 구한 것과 같은 값의 데이터로 그려진 것을 확인할 수 있다.



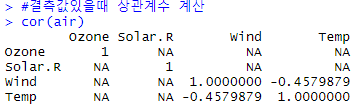


* “Temp”에 대한 상자그림이다. 기초통계량을 구한 것과 같은 값의 데이터로 그려진 것을 확인할 수 있다.

**-상관관계**

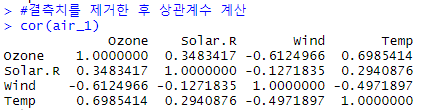
상관 분석은 두 연속 변수가 서로 관련이 있는지 검정하는 통계분석이다. 4개의 변수들을 사용해 각각에 대해 분석을 해볼 것이다.

첫 번째로, 결측치가 있는 데이터(air)를 가지고 상관분석을 해 보았다.

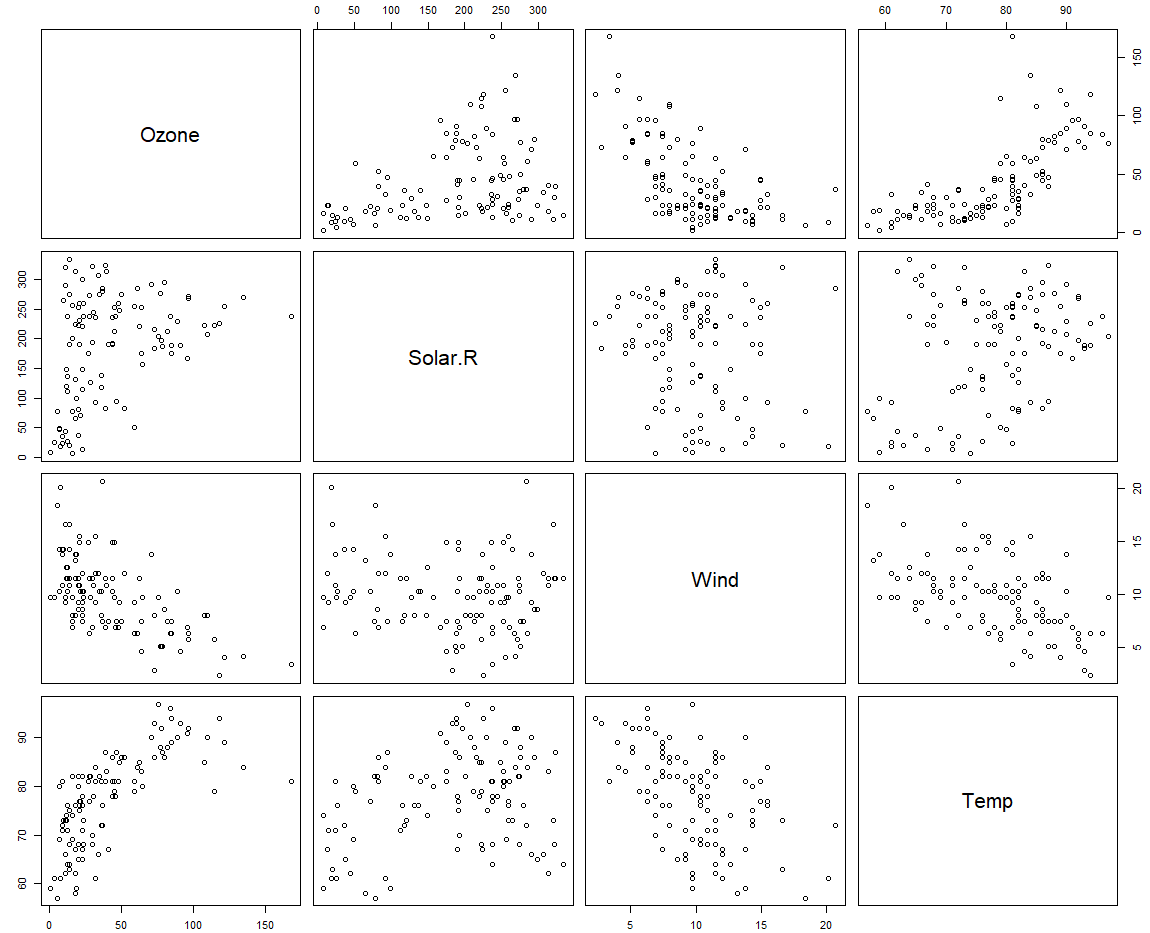


* Na가 포함되어있어 상관계수가 계산되지 않고, NA값으로 나온 것을 볼 수 있다.

두 번째로, 결측치를 제거한 데이터(air\_1)을 가지고 상관분석을 해보고, 그래프를 그려보았다.



* NA값이 제거된 데이터인 4개의 변수들의 상관계수이 계산이 되어 나타난 것을 볼 수 있다.



* 상관계수를 나타낸 그래프를 그린 것이다. 하지만 수치가 써있지 않아 각각의 변수의 상관관계가 어떠한지 뚜렷하게 알 수 없었고, 그래프를 보며 분포만 볼 수 있었다.

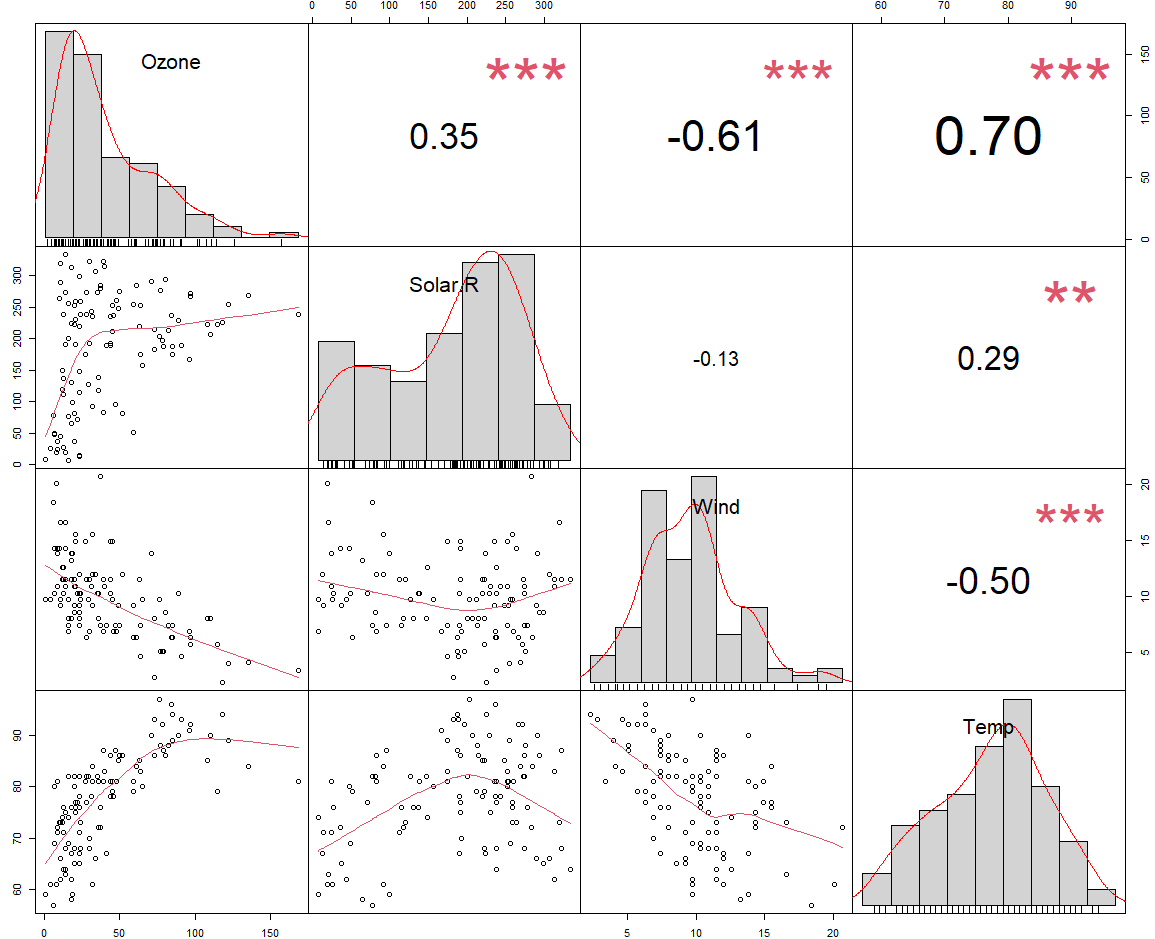
**Ⅲ. 결론**

**ⅰ. 데이터 분석 결과 및 해석**

**-상관분석 결과**

산점도를 가지고 변수들에 대해 상관분석을 해보았다.





* 그래프를 확인한 결과, “Ozone”과 “Temp”가 0.70으로 강한 양적 선형관계라는 사실을 알 수 있고, “Ozone”과 “Solar.R”은 0.35로 뚜렷한 양적 선형관계라고 할 수 있다. 또한 “Ozone”과 “Wind”는 -0.61, “Wind”와 “Temp”는 -0.50으로 뚜렷한 음적 선형관계라고 할 수 있다.

**ⅰⅰ. 새로 알게 된 사실**

airquality데이터를 가지고 기초통계량과 상관관계, 산점도까지 살펴보았다. 기본 데이터 셋을 사용하였지만 코드를 사용해서 그래프를 그리는 과정에서 헷갈리는 것이 많았다. 수업에서 사용했던 코드들을 사용해서 나타내보았더니 금방 적응이 되었던 것 같다. 산점도에서 마이너스가 붙어있으면 선형관계가 형성되지 않는 줄만 알았다. 하지만 그래프도 음적/양적 상관관계가 존재한다는 사실을 새로 알 게 되었다. 또한 기초통계량을 구할 때, summary함수를 이용하여 한번에 편리하게 볼 수 있었고, 상자그림을 해석하면서 summary함수를 이용하면 상자그림의 수치를 정확히 알 수 있구나 라는 사실을 또 한번 깨달았다. 데이터에 NA값이 존재할 경우 NA값을 확인하는 방법, 지우는 방법 등을 다시금 공부할 수 있었다. 마지막으로 혼자 데이터를 찾고, 주제를 찾아보면서 여러 코드를 사용하여 분석을 해보았다. 흥미롭고 유익했었던 시간 이였던 것 같다.

**Ⅳ. 출처**

<https://blog.naver.com/mhy563/222053388603>

<https://rfriend.tistory.com/83>