Module 2 R Practice Assignment: Air Quality Dataset

Trang Tran

CPS, Northeastern University

ALY6010 | Probability Theory and Introductory Statistics

Patrick McQuillan

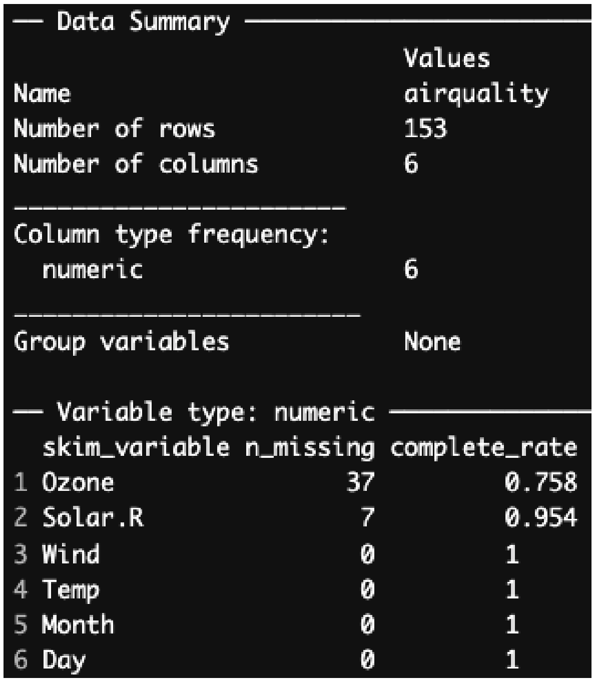
Mar 04, 2023

Introduction

The provided “airquality” dataset consists of 153 rows and 6 columns (variables): Ozone (mean Ozone concentration), Solar.R (Solar radiation), Wind (average wind speed), Temp (maximum daily temperature), Month (Month of observation), and Day of the month.

Descriptive Statistics

Table 1 provides a data summary by using the [skim] function. There are 37 and 7 missing values in the Ozone and Solar columns respectively. All 6 variables are in the numeric class. Therefore, I changed the Month and Date variables into a character type, and they should be factor variables as well.



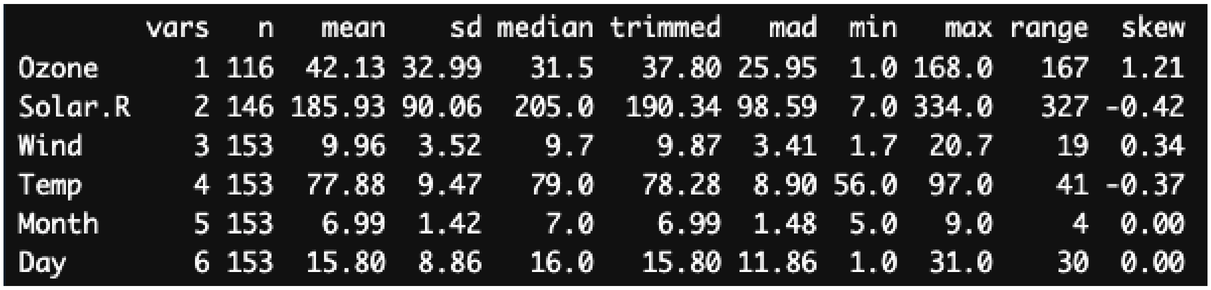
```
── Data Summary ────────────────────────
                           Values
Name                       airquality
Number of rows             153
Number of columns          6
_______________________
Column type frequency:
  numeric                  6
________________________
Group variables            None

── Variable type: numeric ──────────────
  skim_variable n_missing complete_rate
1 Ozone                37         0.758
2 Solar.R               7         0.954
3 Wind                  0         1
4 Temp                  0         1
5 Month                 0         1
6 Day                   0         1
```

Table 1: An overview of dataset using [skim] function.

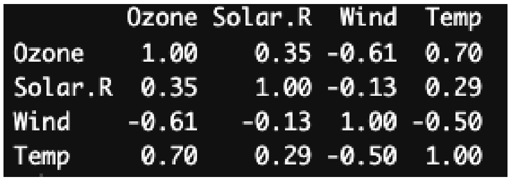
Table 2 below presents another data overview by using the [describe] function, including all dataset parameters. We can see the skewness of the Ozone variable is 1.21 proving that it is significantly positively skewed.

I dropped NAs values in the dataset and then checked the correlation between 4 numeric variables. Also, I grouped the data by Month to see the number of observations after cleaning. (Table 3 & 4)



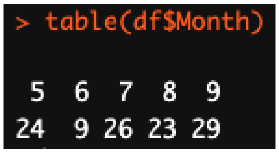
```
        vars   n   mean    sd median trimmed   mad  min   max range  skew
Ozone      1 116  42.13 32.99   31.5   37.80 25.95  1.0 168.0   167  1.21
Solar.R    2 146 185.93 90.06  205.0  190.34 98.59  7.0 334.0   327 -0.42
Wind       3 153   9.96  3.52    9.7    9.87  3.41  1.7  20.7    19  0.34
Temp       4 153  77.88  9.47   79.0   78.28  8.90 56.0  97.0    41 -0.37
Month      5 153   6.99  1.42    7.0    6.99  1.48  5.0   9.0     4  0.00
Day        6 153  15.80  8.86   16.0   15.80 11.86  1.0  31.0    30  0.00
```

Table 2: Another data summary using [describe] function.



```
        Ozone Solar.R  Wind  Temp
Ozone    1.00    0.35 -0.61  0.70
Solar.R  0.35    1.00 -0.13  0.29
Wind    -0.61   -0.13  1.00 -0.50
Temp     0.70    0.29 -0.50  1.00
```

Table 3: Correlation between variables

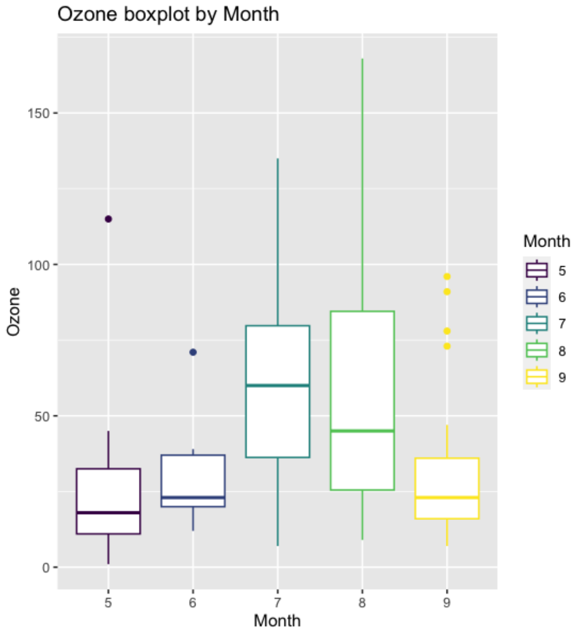


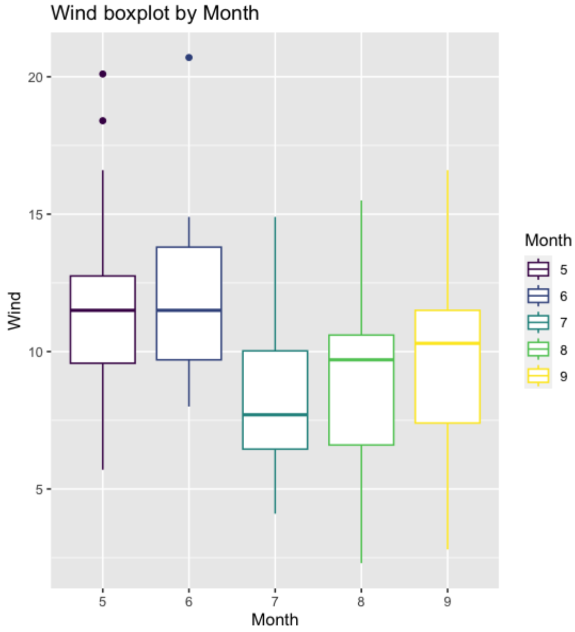
```
> table(df$Month)

 5  6  7  8  9
24  9 26 23 29
```

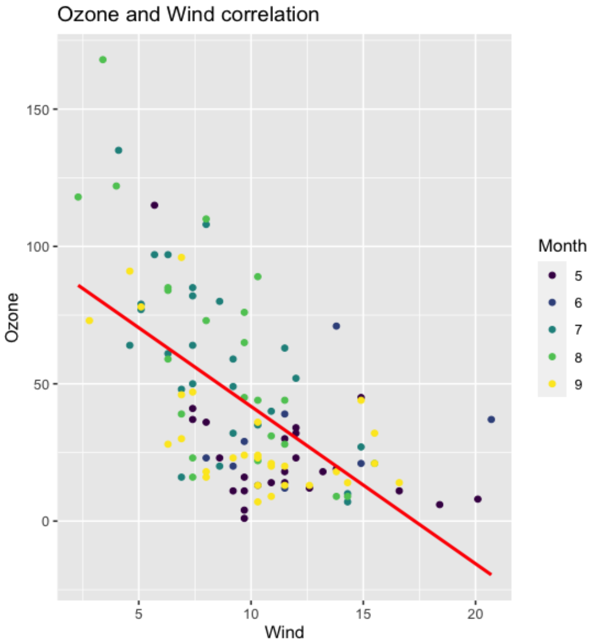
Table 4: Data grouped by Month after cleaning

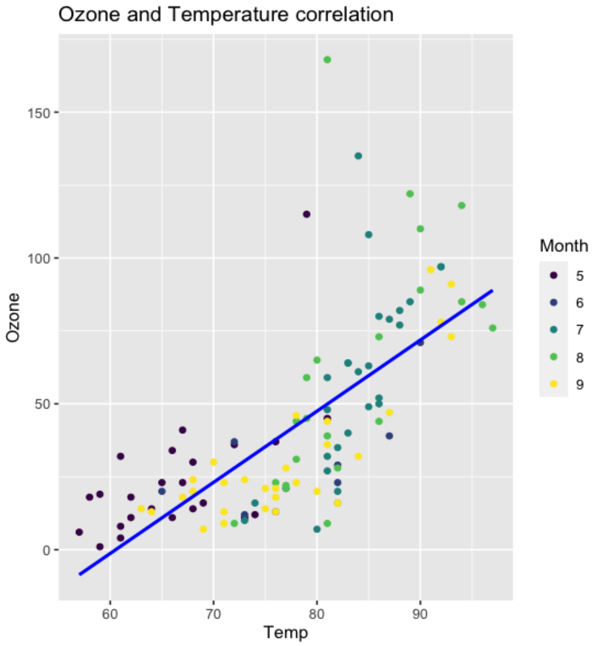
Visualizations and Analysis


Ozone boxplot by Month


Wind boxplot by Month

The ozone boxplot by month shows several outliers mostly in September, and the largest IQR in August. The wind boxplot by month illustrates the close similarity of IQR in all 5 months and the lowest median wind value in July.


Ozone and Wind correlation


Ozone and Temperature correlation

There is a strong negative correlation between Ozone concentration and average wind speed, meanwhile, the Ozone index has a strong positive relationship with the daily temperature over these months.