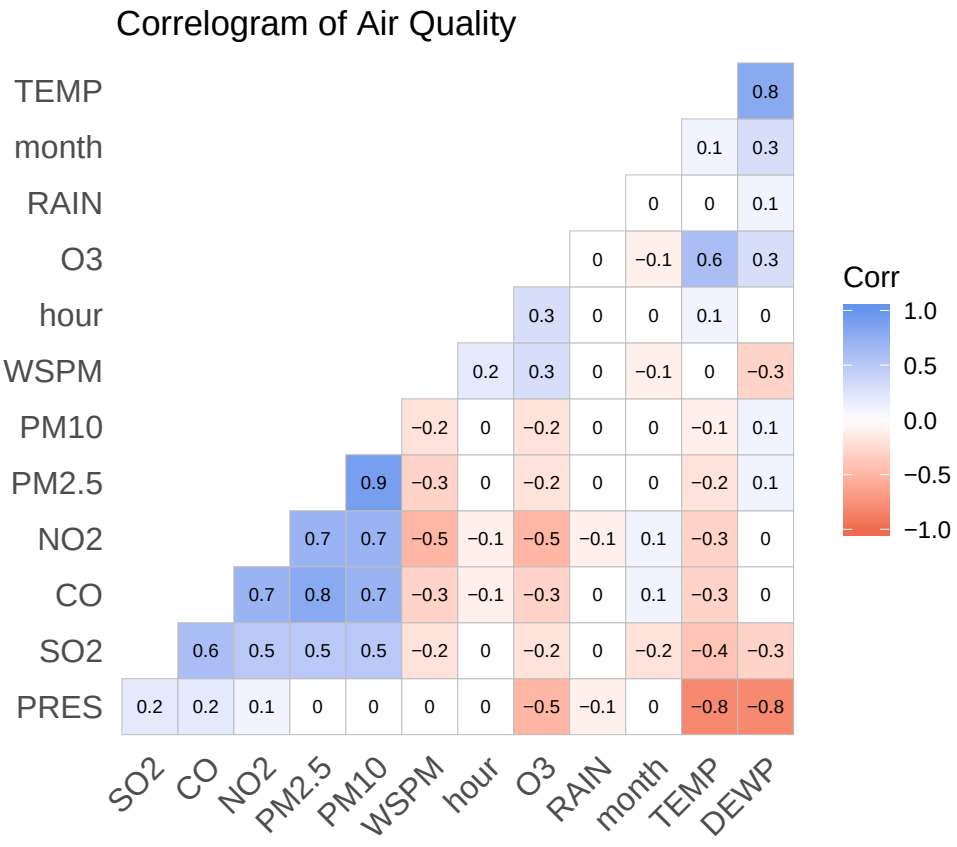
Data Visualization

Meher Ivatury

10/6/2020

Correlation Plot

- This plot using Pearson correlation coefficient to display the relationship between each pair of variables. *Red* refers to negative correlation and *blue* refers to positive correlation. The lighter the color is, the less related they are and vice versa. For instance, `NO2` is highly related to `CO` and `PM2.5` since the correlation coefficient is 0.8, very close to 1, for both of them. The -0.5 between `NO2` and `WSPM` stands for negative correlation. Higher wind speed always brings lower NO2. In terms of the characteristics of Pearson coefficient, we removed `year`, `day` as well as `wd` from this plot.

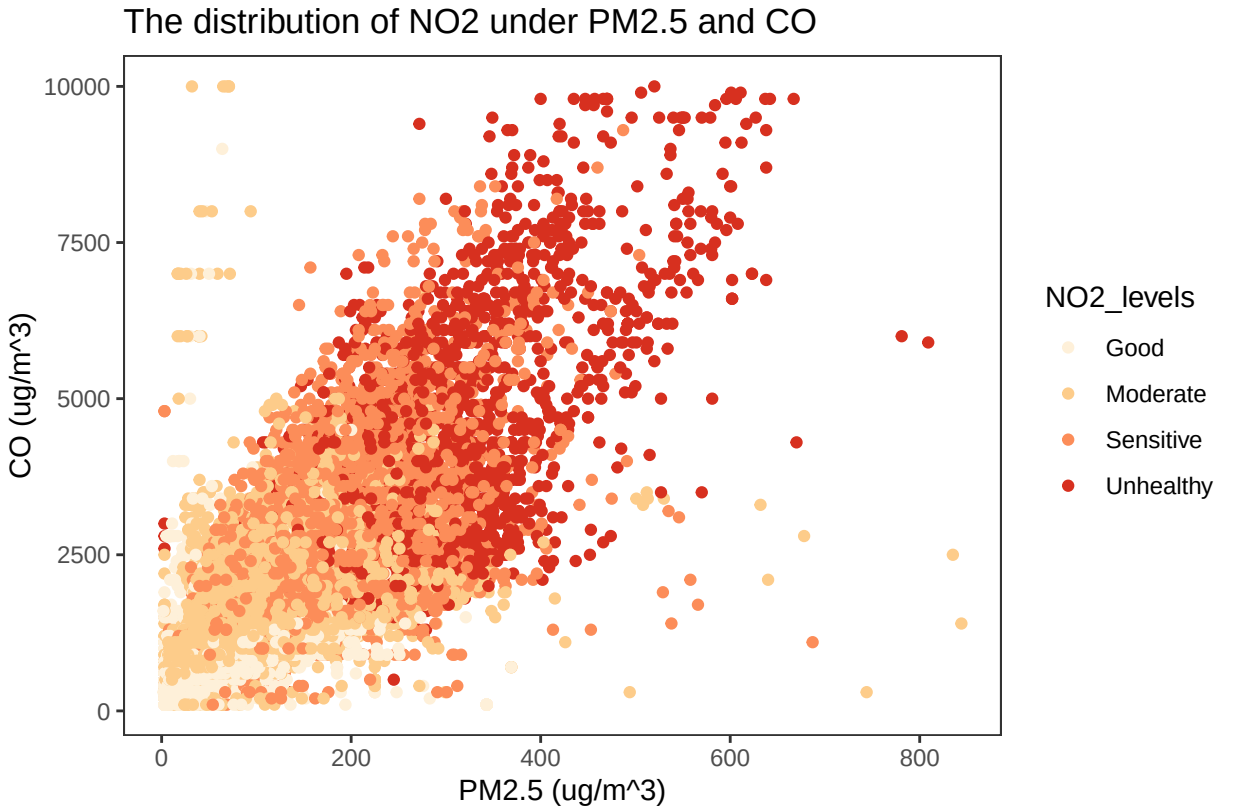


The distribution of NO2 under different conditions

- We first divide the values of NO2 into 4 levels, known as `Good` (values less than or equal to 40), `Moderate` (values between 40 and 80), `Sensitive` (values between 80 and 120) and `Unhealthy` (values greater than 120). This plot display the distribution of `NO2` under different combination of `PM2.5` and `CO`. The various color of points represent each level of `NO2`, lighter color stands for better air quality and darker color stands for worse air quality. It is easy to tell that higher `PM2.5` and higher `CO` are more likely to

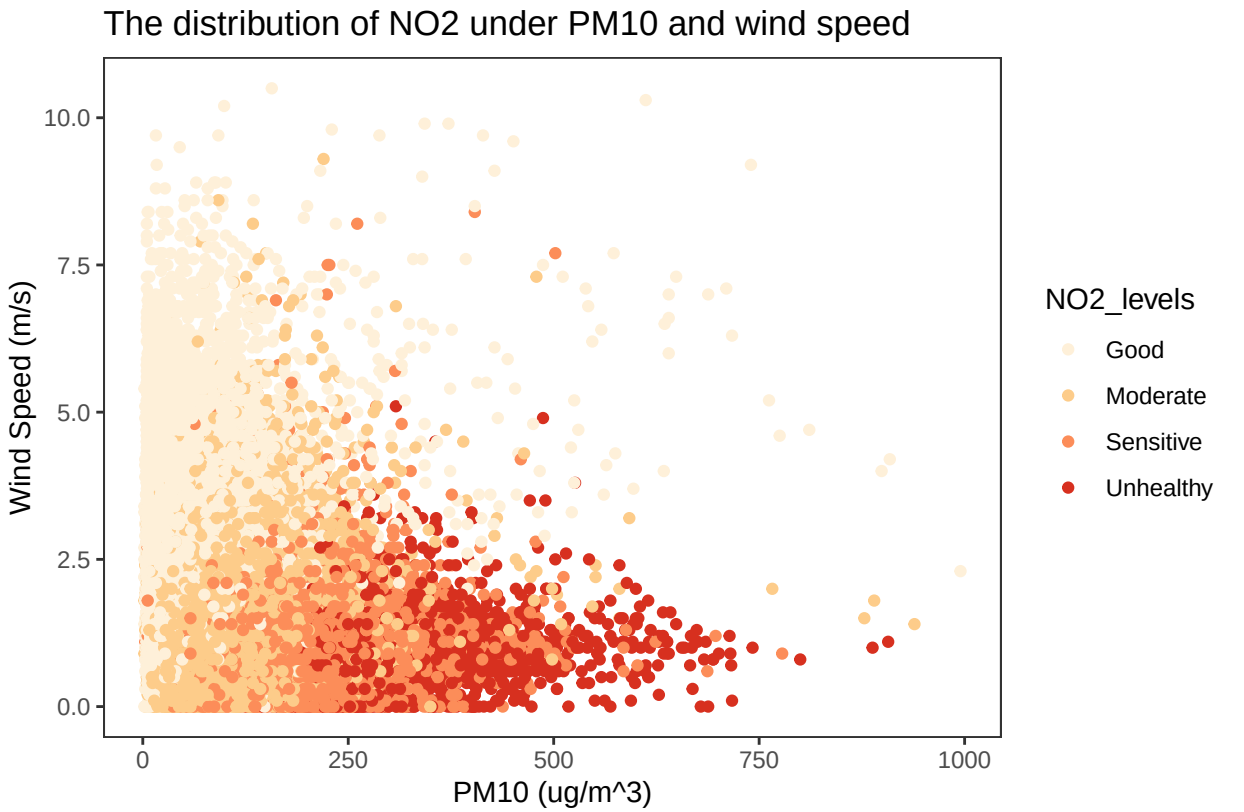
result in *unhealthy air quality*, the highest level of `NO2`.

- As a matter of fact, NO2, CO and PM2.5 are the results of burning of fossil fuels and biomass, and we believe that's why they have positive correlations as shown in the plot.



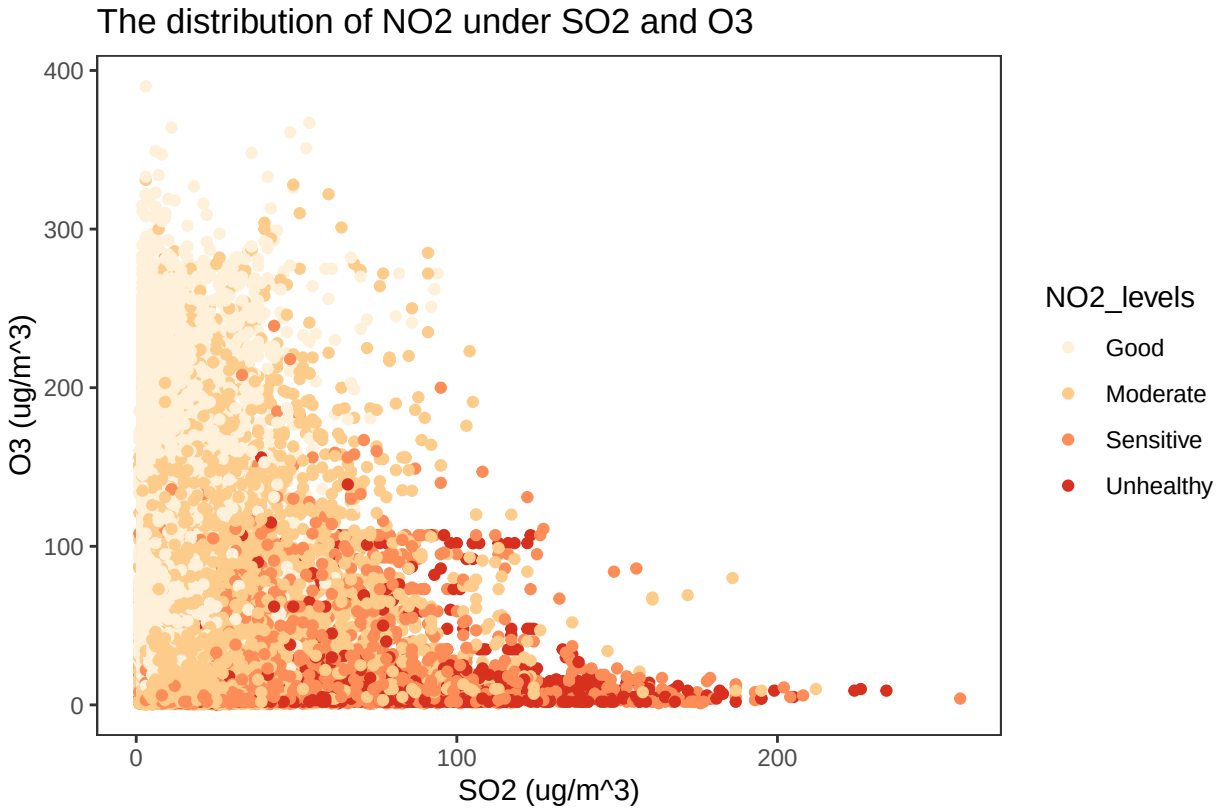
Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

- The graph below reveals the relationship among `NO2`, `PM10` and `WSPM` (wind speed). The majority of darker points are concentrate on lower right of the graph. The highest level of `NO2` is the consequence of high `PM10` and low `WSPM` (wind speed).
- Since PM10 and NO2 have the same source, we believe that is the reason they are positively correlated. However, it's generally known that when the wind is strong, it will blow away all the pollutions and clear the air. In this way, NO2 and wind speed should be negatively correlated. The stronger the wind, the lower the NO2.



Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

- The distribution of `NO2` under the combination of `SO2` and `O3` is shown below. The lighter colored points are concentrate on the upper left part. The relationship between `O3` and `NO2` is negative, that is to say, Higher `O3` results in lower `NO2`. However, higher `SO2` results in higher `NO2` and in other words, `SO2` and `NO2` are positive related.
- According to what we found about the characteristics of NO2, it will break up into NO and Oxygen atom when exposed to ultraviolet light and then the Oxygen atom will combine with oxygen in the air into O3. In this way, as NO2 decompose and decrease, O3 will increase. We would like to take this as the reason of the negative correlation between NO2 and O3.

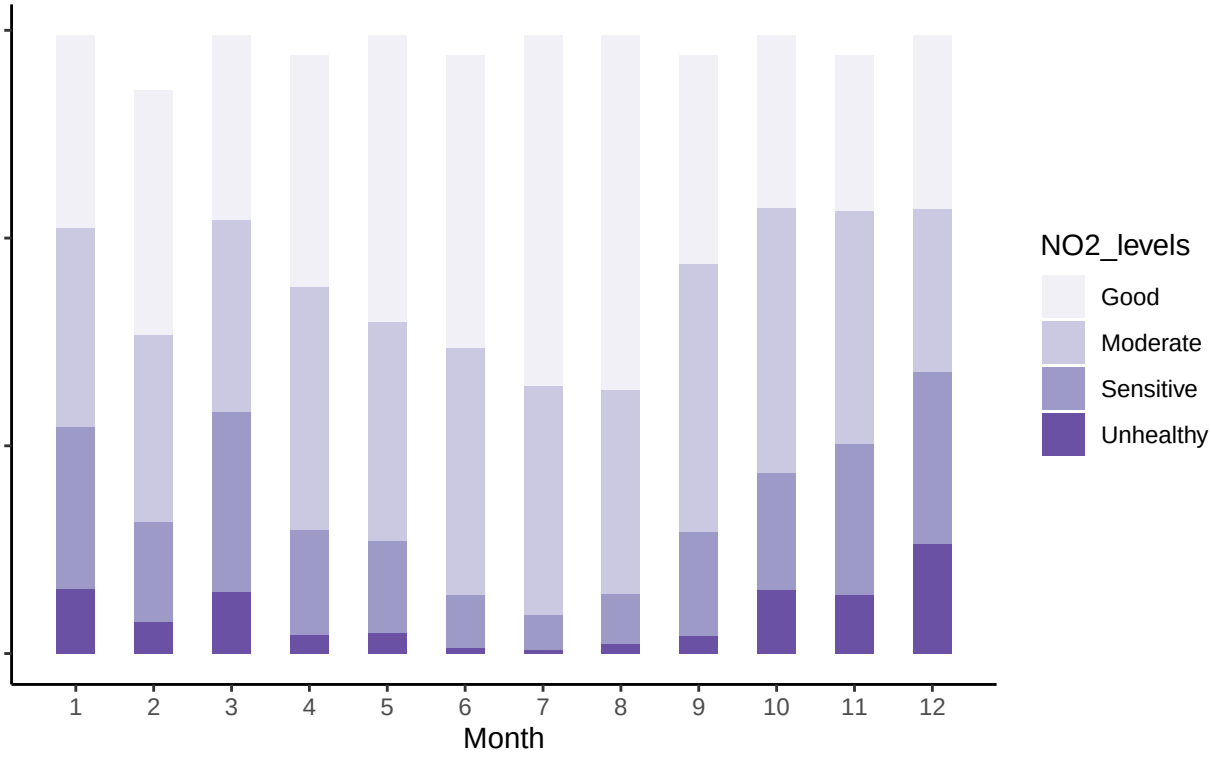


Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

Distribution of NO2 in each month and season

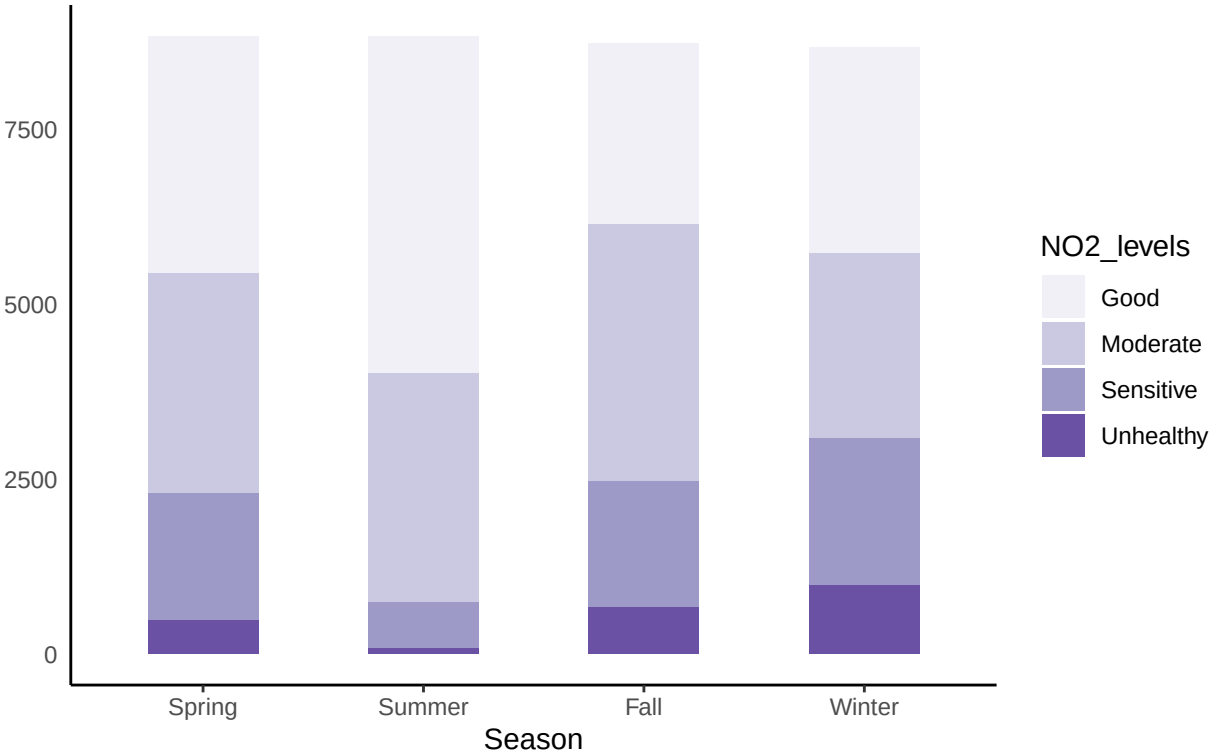
- The two graphs below demonstrate the distribution of the 4 levels of `NO2` in each month and in each season. For instance, the distribution of *unhealthy* has a greater proportion in December than in July. The distribution of *sensitive* and *unhealthy* take up more percentage in fall and winter than in summer, that is to say, the air quality is worse in fall and winter. And obviously, *good air quality* takes up to 65% days in July and August.
- Since Beijing is located at the northen hemisphere and it has the temperate monsoon climate, it has the climate pattern that winter cold with less preciptation and summer hot with more precipitation. On the one hand, winter in Beijing is cold and people always use heating in this season, which is powered by coal burning and it also result in higher NO2. This could be one of the reasons that NO2 is higher in winter than in summer. On the other hand, there is more sunlight in summer and it could help in decomposing NO2 to a great extent, which we believe is also the reason that NO2 is lower in summer.

The distribution of NO2 in each month



Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

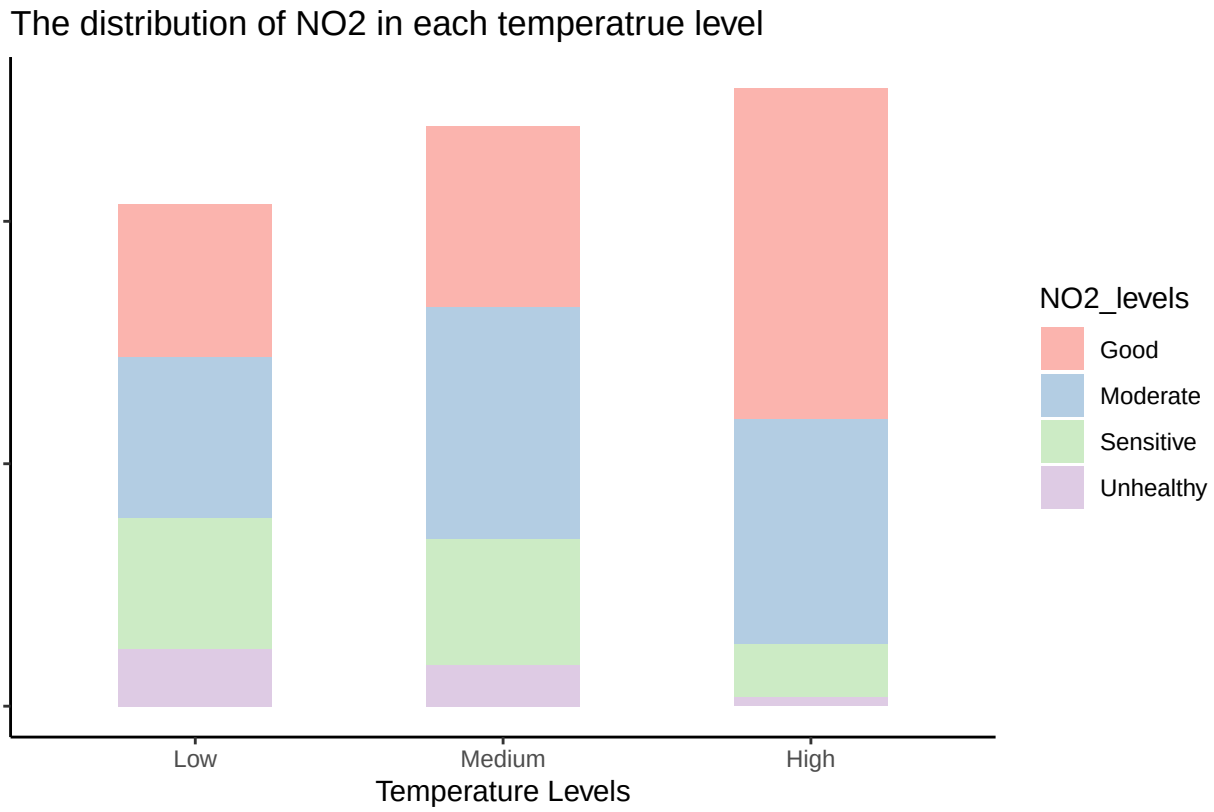
The distribution of NO2 in each season



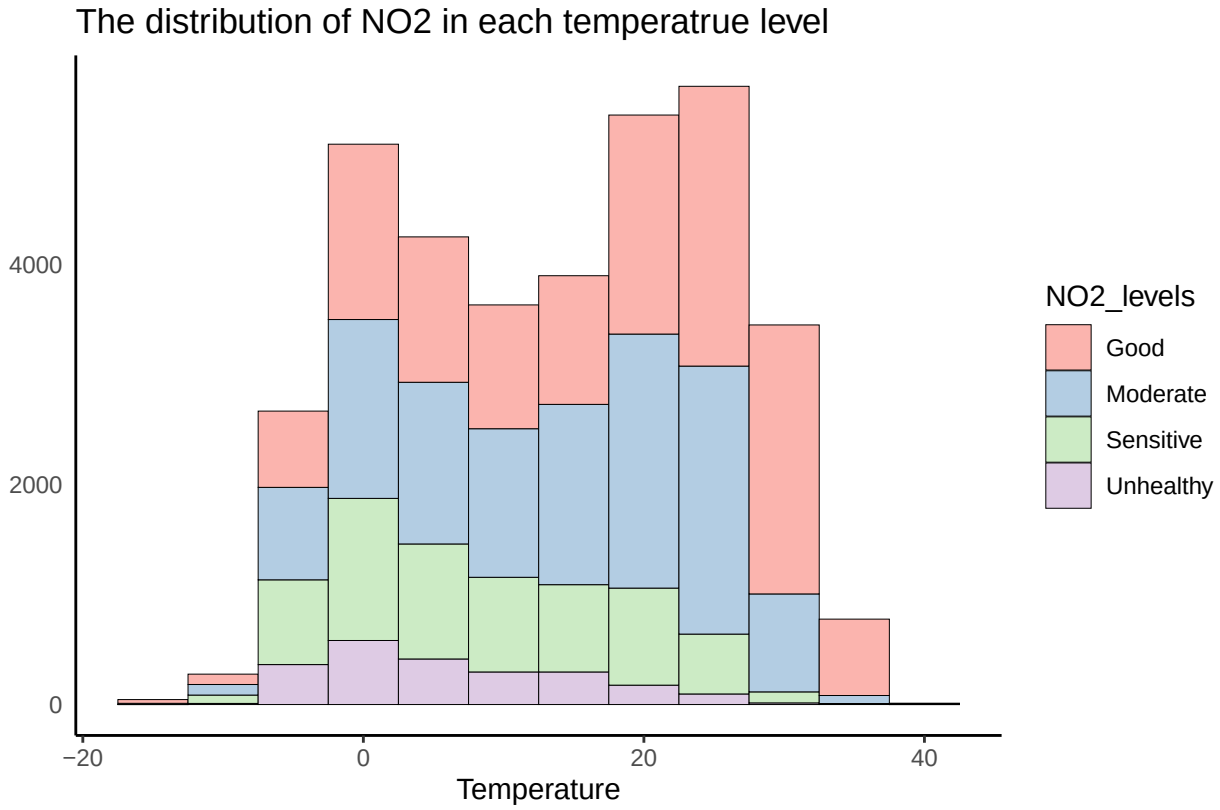
Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

Distribution of NO2 under different temperature level

- We divided temperature into three levels, known as *low temperature* (less than or equal to 5°C), *medium temperature* (greater than 5°C and less than 20°C) as well as *high temperature* (greater than 20°C). The following two graphs reveal the distribution of `NO2` in different `temperatures`. The percentage of *sensitive* and *unhealthy* is greater in *low temperature* than in *high temperature* significantly. We might assume that `temperature` is highly related to the levels of `NO2`.



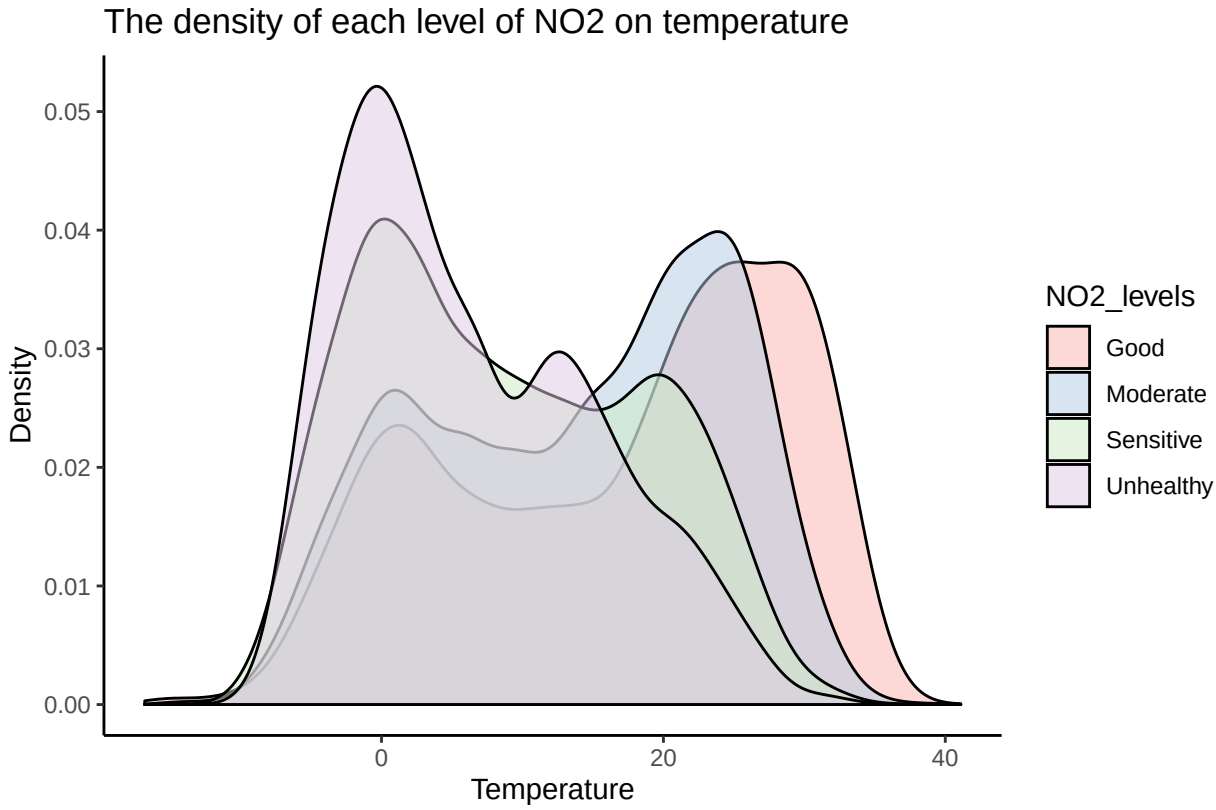
Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data



Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

Density Plot

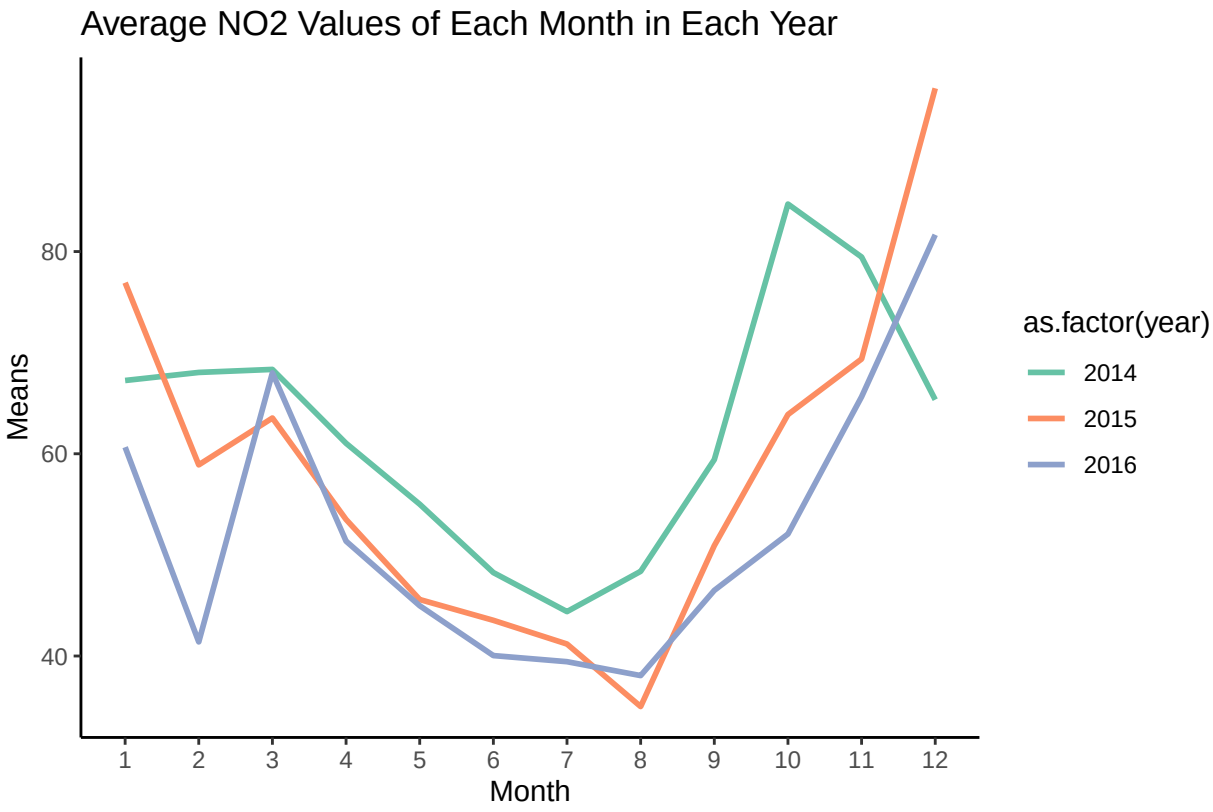
- This is a density plot of each level of `NO2` across `temperatures`. Each color represents the distribution of each level. From this plot we can easily tell the distribution of each level and without regard to the sample size of them. For instance, the highest point of the density curve of *unhealthy air quality* presents at 0°C, that is to say, about 6% of observations that is defined as *unhealthy* shown near 0°C. Besides, the distribution of *good air quality* has a wide range but more concentrate on higher temperatures.



Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

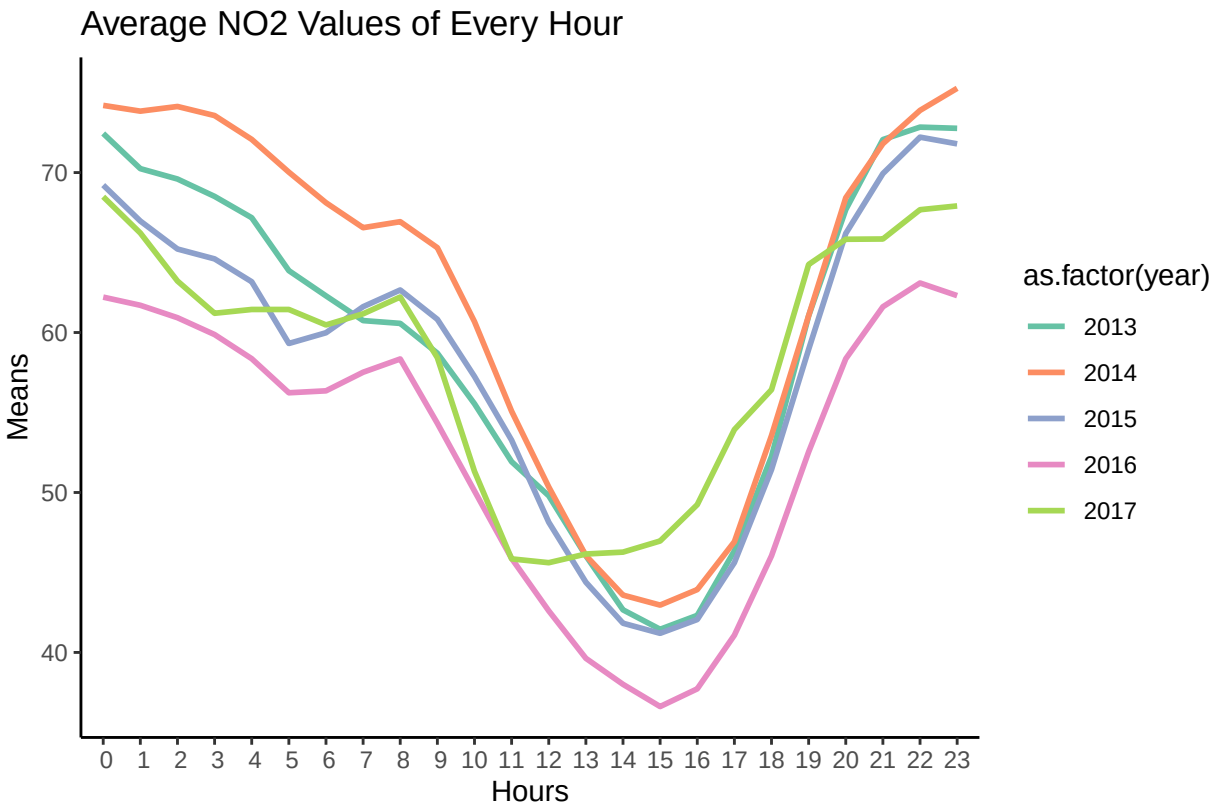
Time Series Plot

- Computing the average `NO2` through each month in each year and exhibiting time series plot below. We can find a distinct pattern over years. Since the data of 2013 and 2017 are incomplete, we only keep the data of 2014, 2015 and 2016. The overall pattern is as same as we mentioned previously, the NO2 values are lower in summer and higher in winter. However, there is a specific feature of this plot, that is the clear decrease in February. Because the Chinese New Year is around this time, the majority of Chinese people would stay at home and spend time with their families. People will have a long vacation and reduce vehicle transportations. We would like to take this as the reason of the decline of NO2 in February.

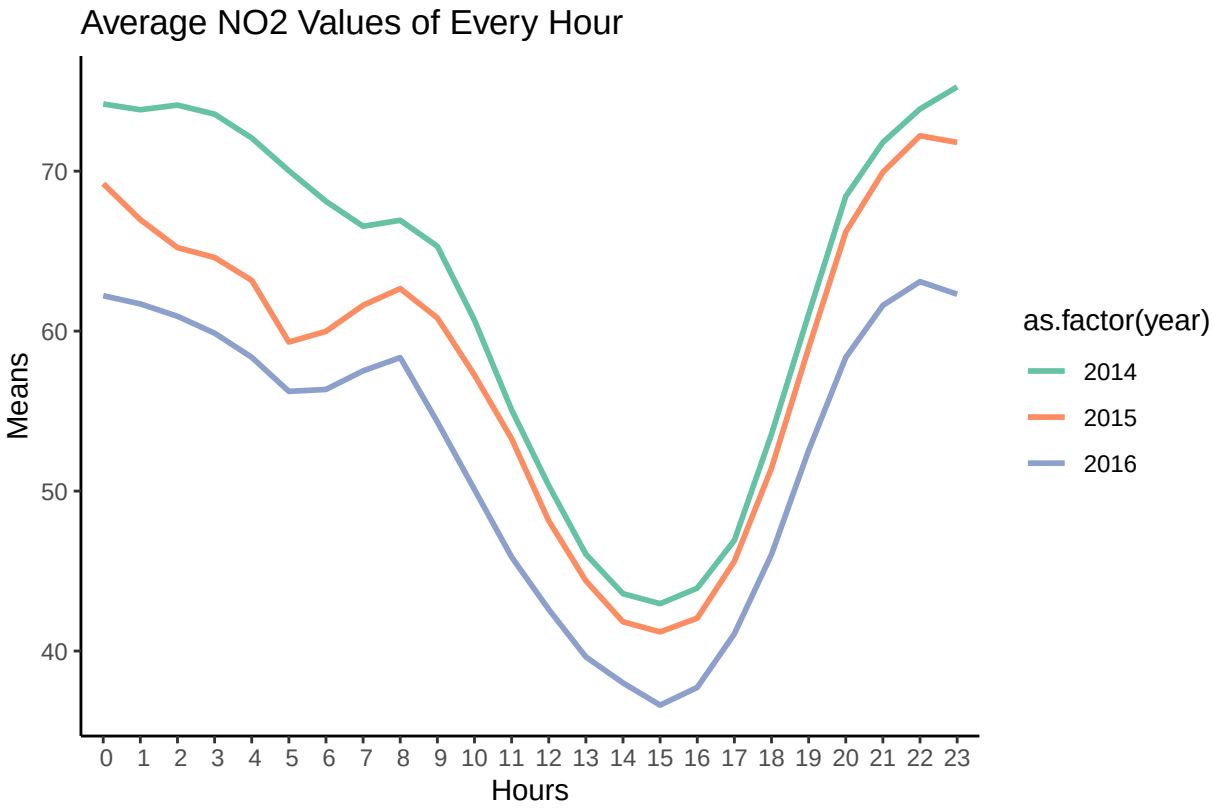


Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

- Then we calculate the average of NO2 values through each hour in every year. The plot is shown below. However, since the data of 2013 begins at March and the data of 2017 only includes Janurary and February, the average NO2 values may be a little biased. For the second plot, we only keep the data of year 2014, 2015 and 2016.
- From this plot, we can find that the mean of NO2 is decrease year by year. It might be the result of the government attaches increasing importance to air quality in recent years.



Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data



Source: https://archive.ics.uci.edu/ml/datasets/Beijing+Multi−Site+Air−Quality+Data

Simple Linear Regression Model

```
##
## Call:
## lm(formula = NO2 ~ as.factor(year) + season + hour + PM2.5 +
##     PM10 + SO2 + CO + O3 + TEMP + PRES + DEWP + RAIN + WSPM,
##     data = airdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -134.822  -11.878   -1.238   10.355  126.204
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.673e+02  2.158e+01  12.388  < 2e-16 ***
## as.factor(year)2014   1.618e+00  3.088e-01   5.238 1.63e-07 ***
## as.factor(year)2015   2.207e+00  3.117e-01   7.080 1.47e-12 ***
## as.factor(year)2016 -1.230e+00  3.154e-01  -3.900 9.65e-05 ***
## as.factor(year)2017 -3.356e+00  6.151e-01  -5.455 4.93e-08 ***
## seasonSummer        -9.143e+00  3.960e-01 -23.087  < 2e-16 ***
## seasonFall          -5.162e+00  3.434e-01 -15.032  < 2e-16 ***
## seasonWinter        -5.143e+00  4.116e-01 -12.495  < 2e-16 ***
## hour                 4.587e-01  1.590e-02  28.855  < 2e-16 ***
## PM2.5                4.264e-02  3.305e-03  12.901  < 2e-16 ***
## PM10                 9.942e-02  2.659e-03  37.390  < 2e-16 ***
## SO2                  1.436e-01  5.941e-03  24.170  < 2e-16 ***
## CO                   6.458e-03  1.553e-04  41.586  < 2e-16 ***
## O3                  -2.554e-01  2.589e-03 -98.645  < 2e-16 ***
## TEMP                 3.677e-01  2.808e-02  13.095  < 2e-16 ***
## PRES                -2.119e-01  2.121e-02  -9.993  < 2e-16 ***
## DEWP                -1.120e-01  2.029e-02  -5.519 3.43e-08 ***
## RAIN                -8.578e-01  1.309e-01  -6.552 5.74e-11 ***
## WSPM                -5.946e+00  9.954e-02 -59.733  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.99 on 35045 degrees of freedom
## Multiple R-squared:  0.7291, Adjusted R-squared:  0.7289
## F-statistic:  5239 on 18 and 35045 DF,  p-value: < 2.2e-16
```