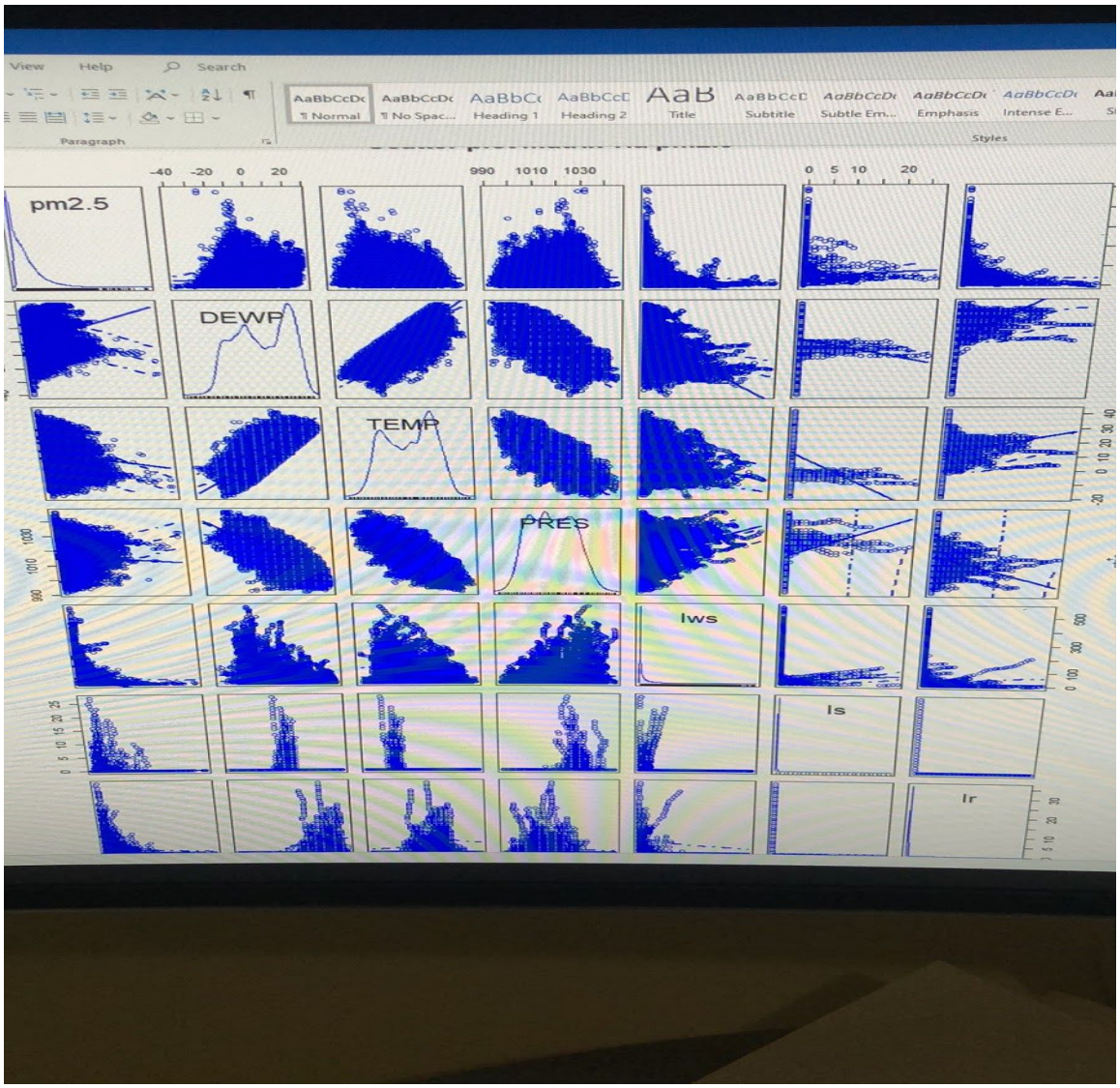
Yifei Chen
STA 141A
10/30/2019

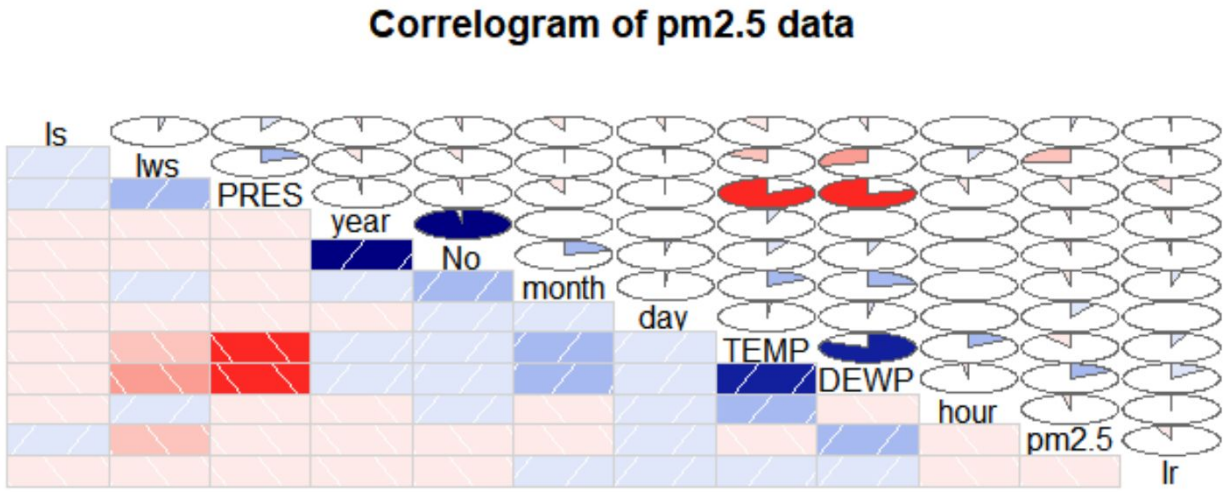
Honor Code: The codes and results derived by using these codes constitute my own work. I have consulted the following resources regarding this assignment: Chao Cheng

1. Provide a graphical summary, with explanations, to describe how the various numeric variables relate to each other.



(I only show the screenshot because it will take me several hours to reload the graph again) According to the plot, we can see that there is a strong positive relationship between Dew Point(DEWP) and temperature(TEMP). And there exists a strong negative relationship between Dew Point(DEWP)/temperature(TEMP) and pressure(PRES). From the scatterplot, we can also see that for Dew Point(DEWP)/temperature(TEMP)/pressure(PRES) vs

Cumulated hours of snow(Is)/Cumulated hours of rain(Ir), most of the dots concentrate in the center. For pm2.5 vs Cumulated hours of snow(Is)/Cumulated hours of rain(Ir)/Cumulated wind speed(Iws), the dots mostly gather at the beginning and decrease rapidly as pm2.5 increases. Except for Iws, Is, Ir, it is hard to find the pattern of pm2.5 with the other variable, such as TEMP, PRES. Therefore, more analyzations are needed.



The correlogram uses pie charts to show the relationship between each variable. Similar to the previous graph, correlogram also shows a strong negative relationship between TEMP/DEWP and PRES, and a strong positive relationship between TEMP and DEWP. From the graph, we can also see a weak negative relationship between Iws and TEMP/DEWP/pm2.5, and weak positive relationships between Iws and PRES, month and TEMP/DEWP, TEMP and hour, and DEWP and pm2.5. The relationships between other variables are relatively weaker.

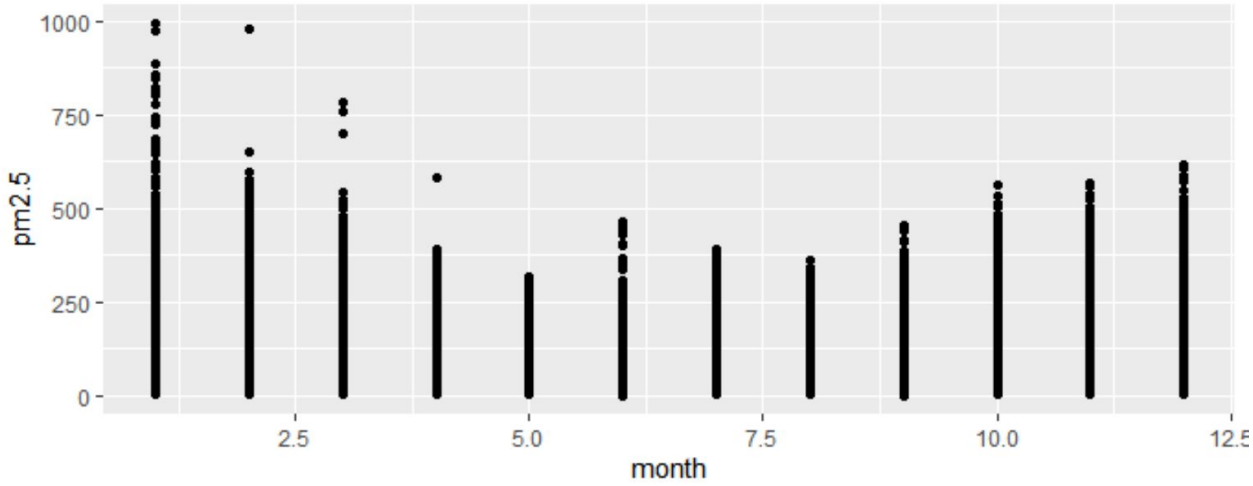
2. Briefly describe if you find any relationship between pm2.5 and month. How do such patterns change across hour and cbwd?

```
> data.dfz = ggplot(data, aes(x=month,y=pm2.5))
> data.lm1 = lm(pm2.5 ~ month, data=data)
> summary(data.lm1)
Call:
lm(formula = pm2.5 ~ month, data = data)

Residuals:
   Min     1Q  Median     3Q    Max
-100.15  -69.30  -26.23   38.42  891.85

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 102.7912     0.9613  106.93  < 2e-16 ***
month        -0.6414     0.1304   -4.92  8.7e-07 ***
```

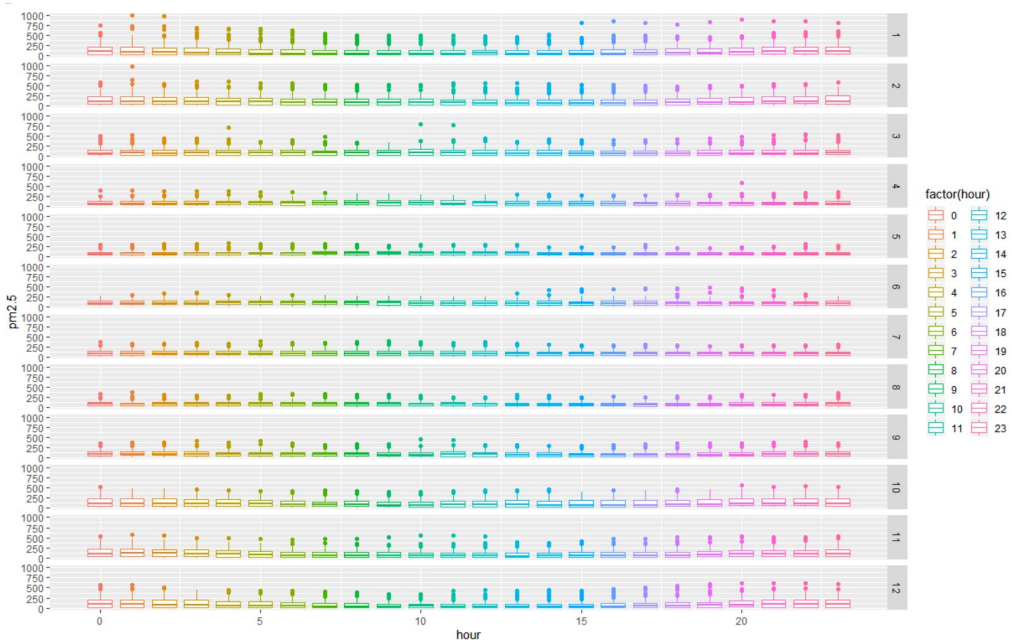
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 92.02 on 41755 degrees of freedom
Multiple R-squared: 0.0005793,	Adjusted R-squared: 0.0005554
F-statistic: 24.2 on 1 and 41755 DF, p-value: 8.703e-07

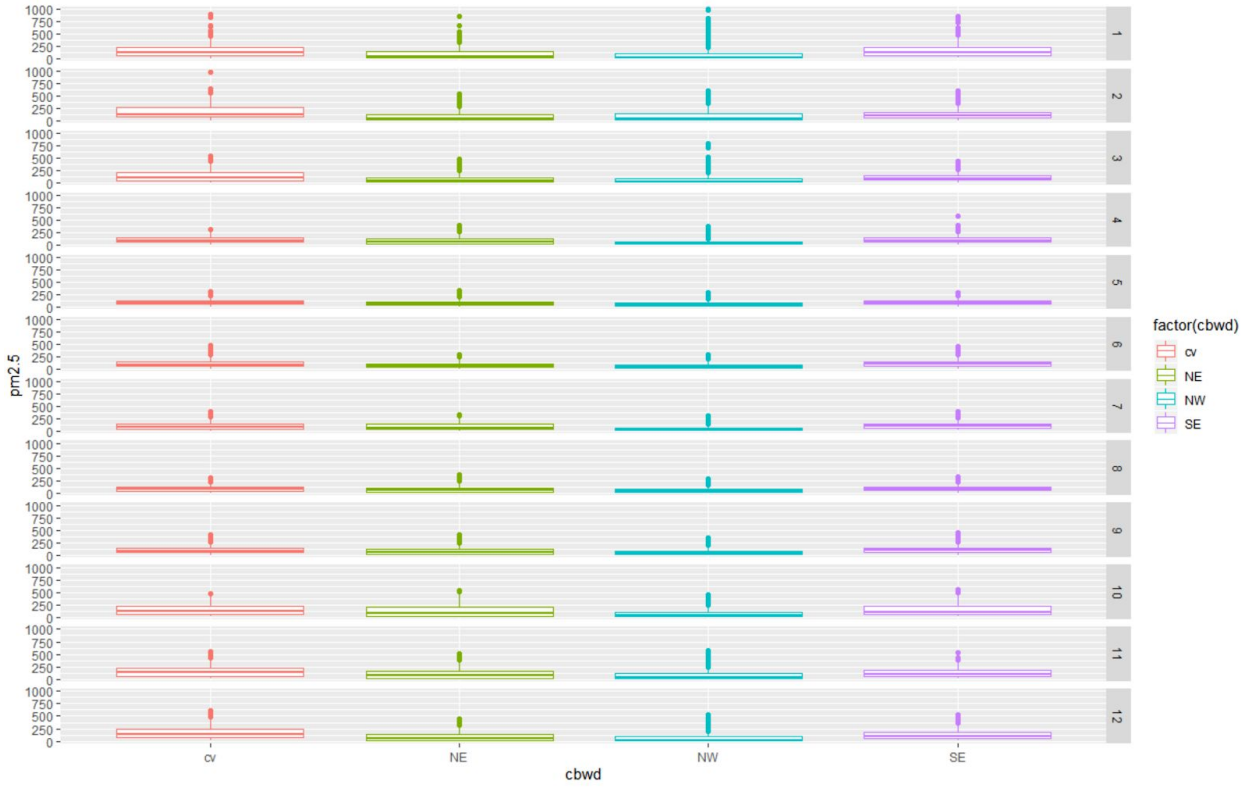
According to the code, if we only compare pm2.5 and month using the summary function, the R-squared is around 0.0006, which is relatively low. It means that the model hardly explains any of the variations in the pm2.5 around its mean. The mean of the dependent variable, which is pm2.5 in here, predicts the dependent variable as well as the regression model.



The scatterplot gives us a better understanding of these two variables. The value of pm2.5 reaches its maximum value in January, which is around 1000, then starts to decrease until the middle of the year, and increases again.

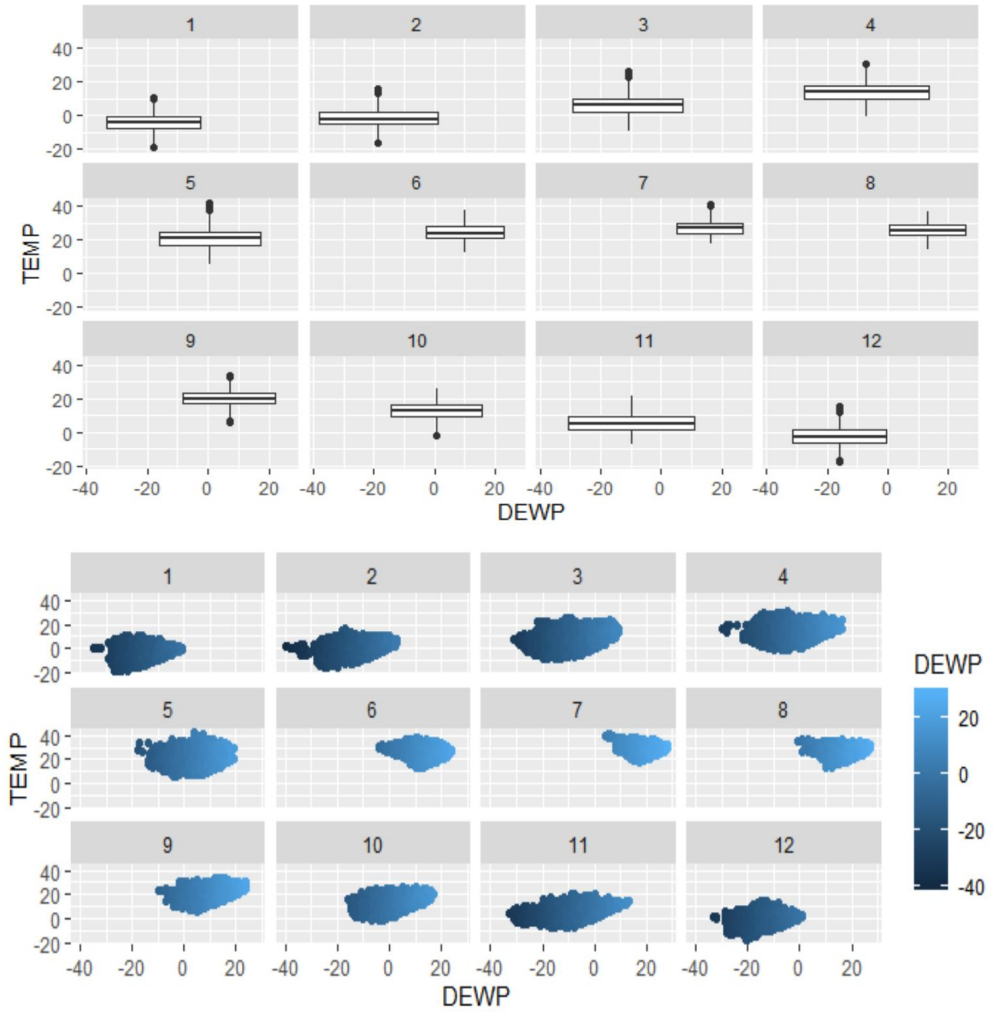


If we add hour as another factor, from the boxplots, in most of the months, the pm2.5 values will decrease until the middle of the day(around 12 pm), then increases. So hour can also be counted as an important factor in predicting the value of pm2.5.

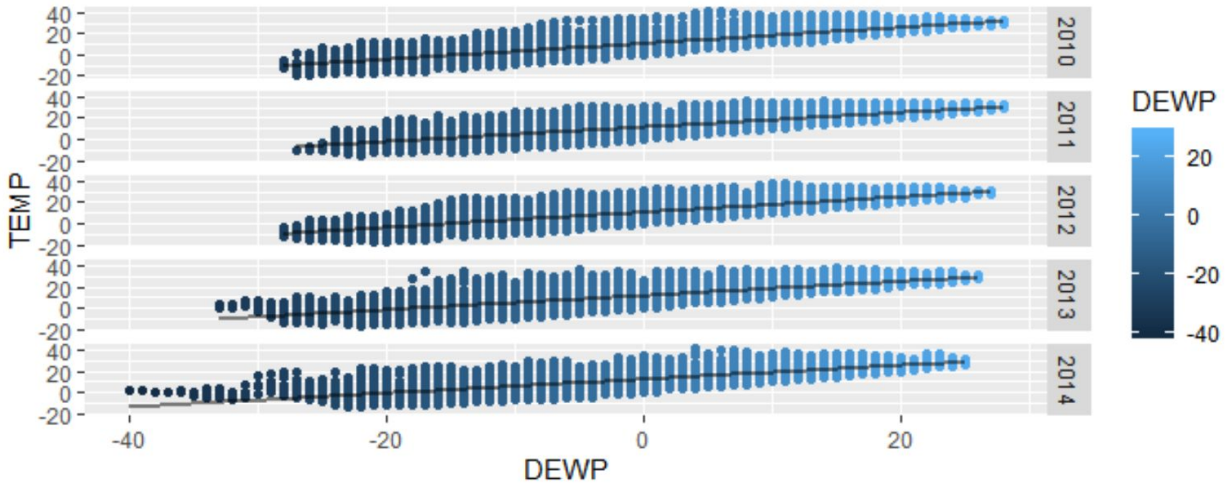


By adding cbwd as another factor, from the boxplots, it also exists some patterns between cbwd and pm2.5 each month. For example, the value of pm2.5 is always a little bit lower when cbwd is NW, and the value will get higher when cbwd is CV.

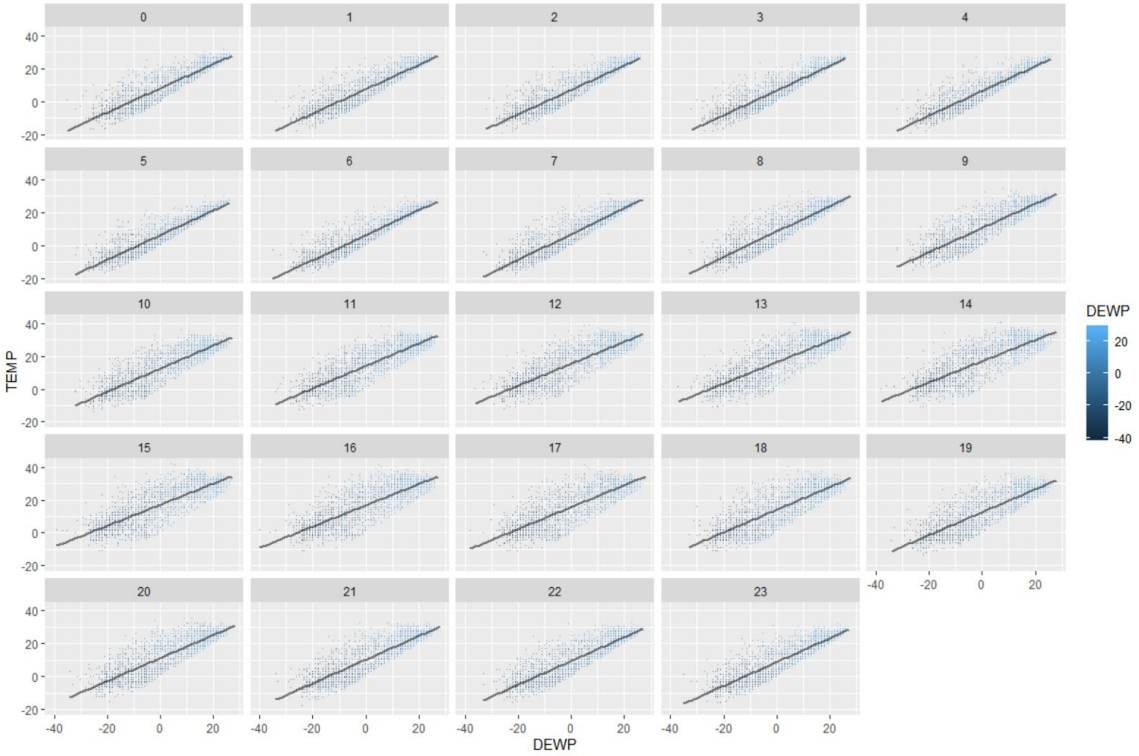
3. Consider the variables DEWP, TEMP, and PRES and describe how the relationships among any pair of these variables vary across strata determined by factors such as year, month and hour.



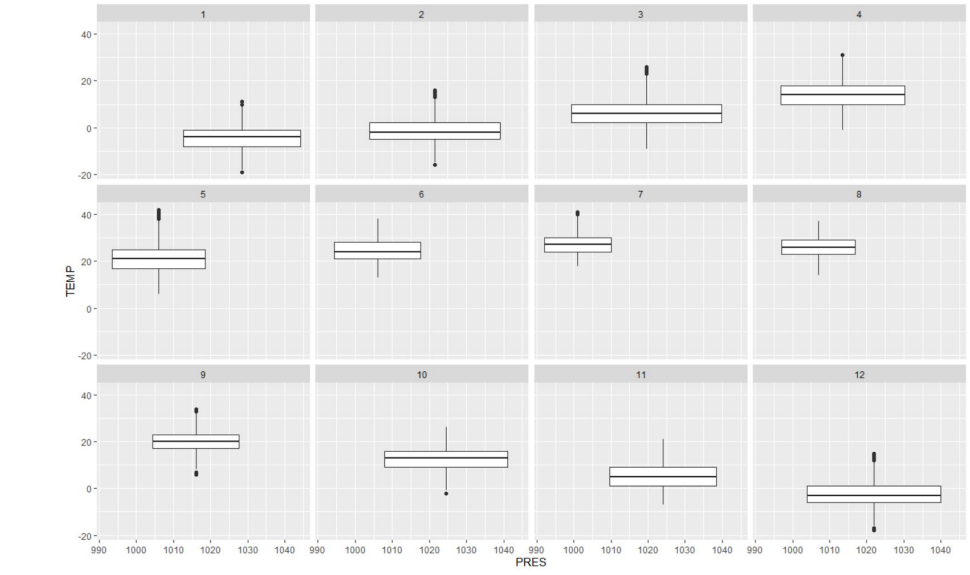
By comparing TEMP and DEWP **in conditional to month** using both scatterplot and boxplot, both TEMP and DEWP increase until the middle of the year(around July), then start to decrease.



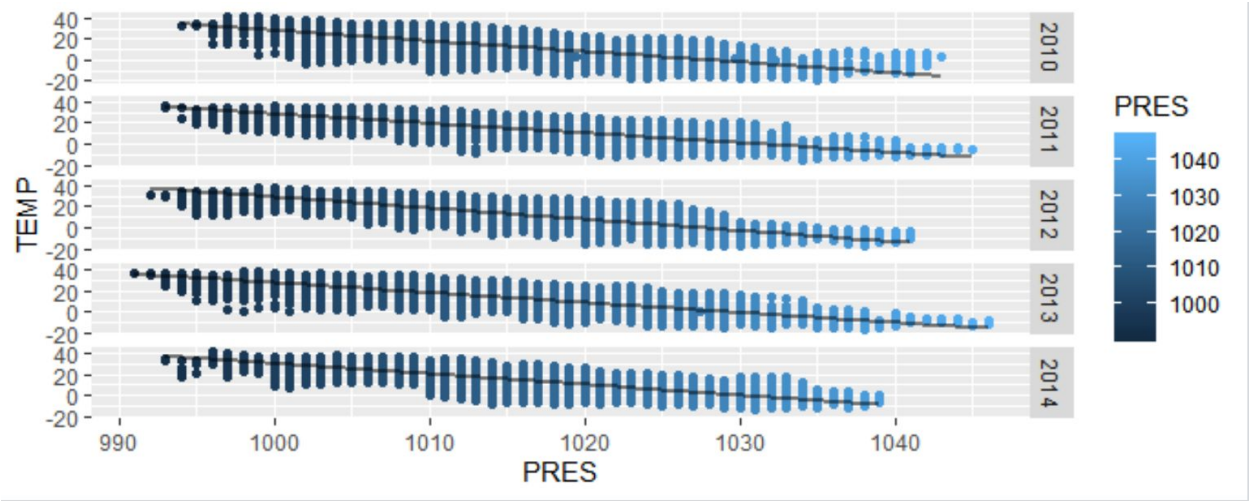
By comparing TEMP and DEWP **in conditional to year** using scatterplot, the pattern of TEMP and DEWP tend to stay the same as the year increase, showing a positive relationship. Therefore, year is not a significant variable in comparing TEMP and DEWP.



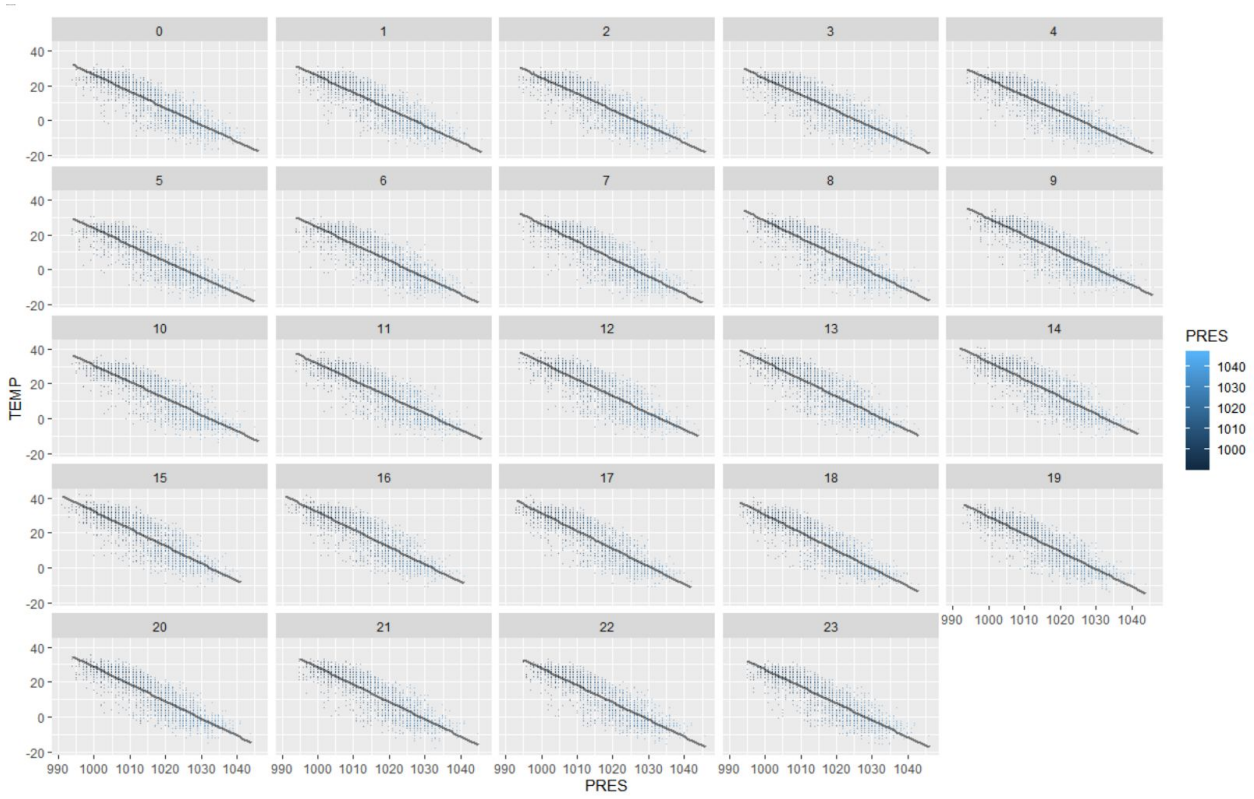
By comparing TEMP and DEWP **in conditional to hour**, although the patterns are similar in each hour, where TEMP and DEWP have a strong positive relationship, the dots seem to be more separated and scattered in the middle of the day(around 10 to 17).



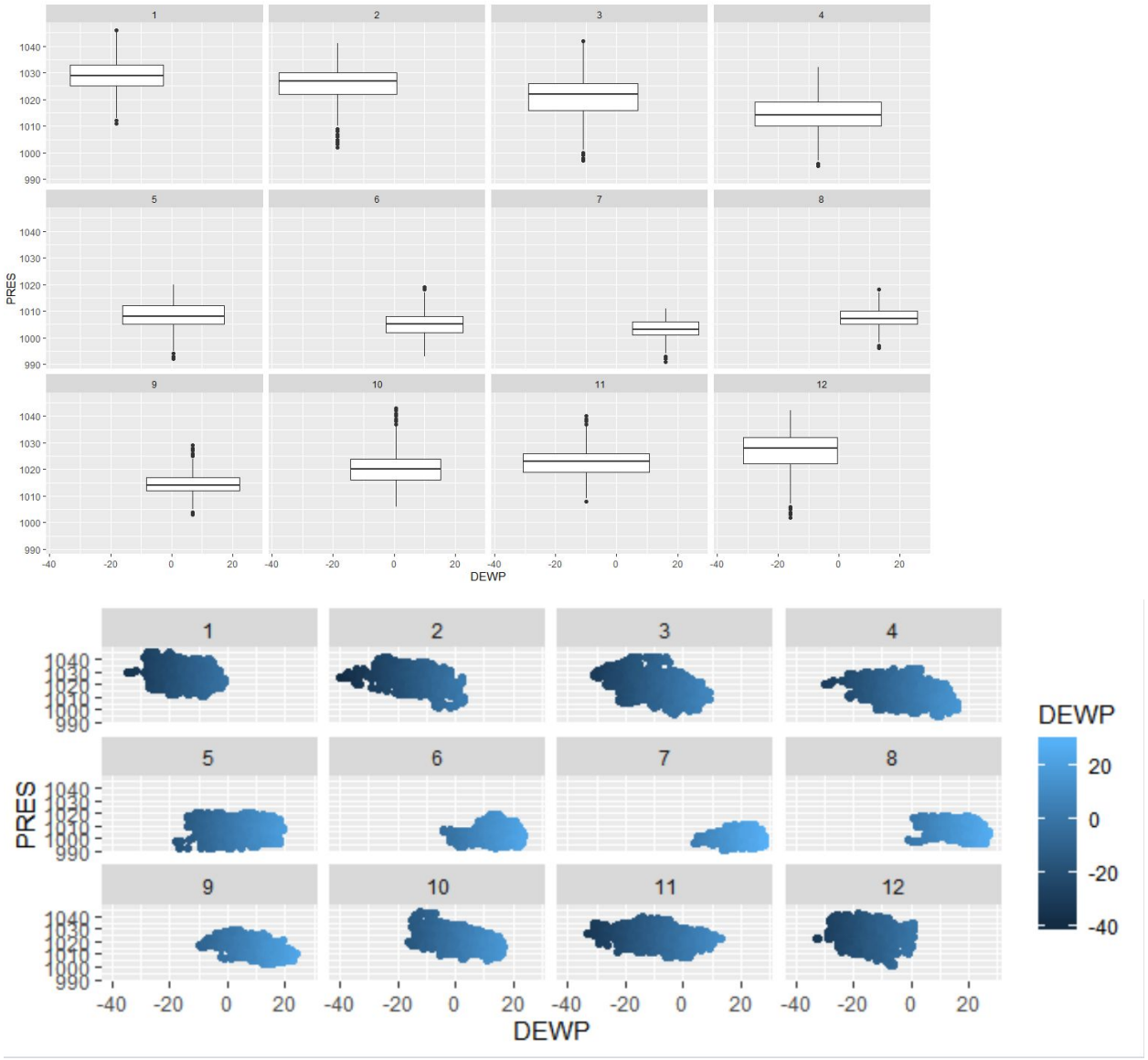
By comparing TEMP and PRES **in conditional to month** using both scatterplot and boxplot, TEMP increases and PRES decreases until the middle of the year(around July), then TEMP starts decreasing while PRES increases.



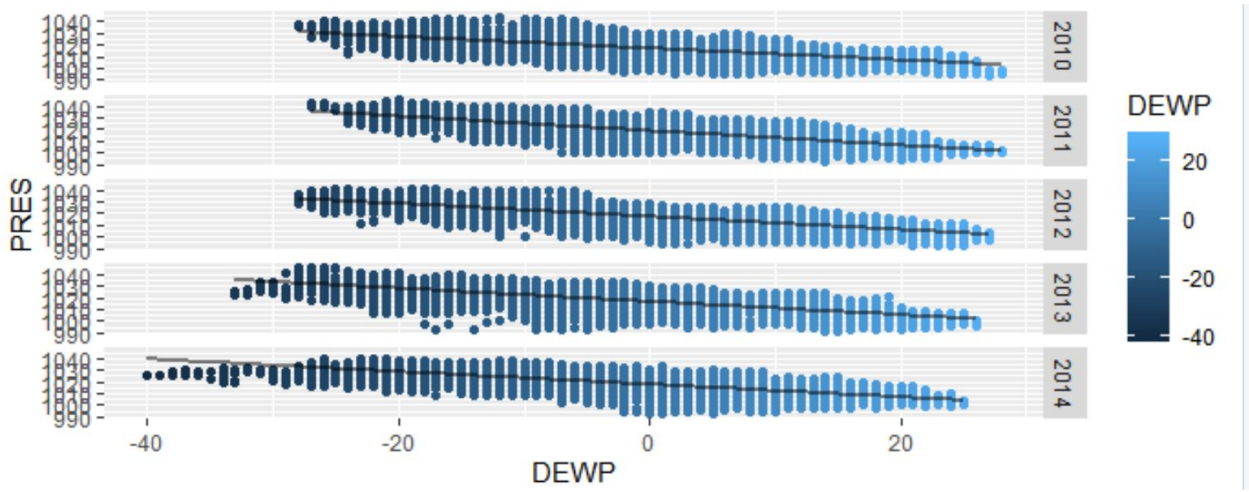
By comparing TEMP and PRES **in conditional to year** using scatterplot, the pattern of TEMP and PRES tend to stay the same as the year increase, showing a strong negative relationship. Therefore, year is not a significant variable in comparing TEMP and PRES.



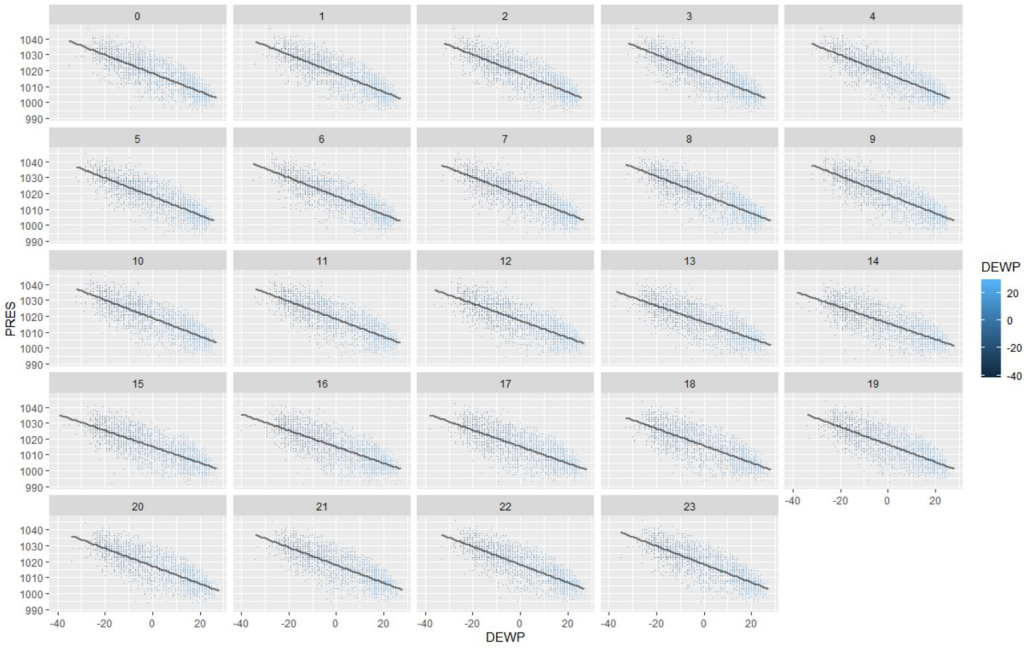
By comparing TEMP and PRES **in conditional to hour**, patterns are similar in each hour, showing a strong negative relationship. So hour is not a significant factor in comparing TEMP and PRES.



By comparing DEWP and PRES **in conditional to month** using both scatterplot and boxplot, DEWP increases and PRES decreases until the middle of the year(around July), then DEWP starts decreasing while PRES increases.



By comparing DEWP and PRES **in conditional to year** using scatterplot, the pattern of DEWP and PRES tends to stay the same as the year increase, showing a strong negative relationship. Therefore, year is not a significant variable in comparing DEWP and PRES.



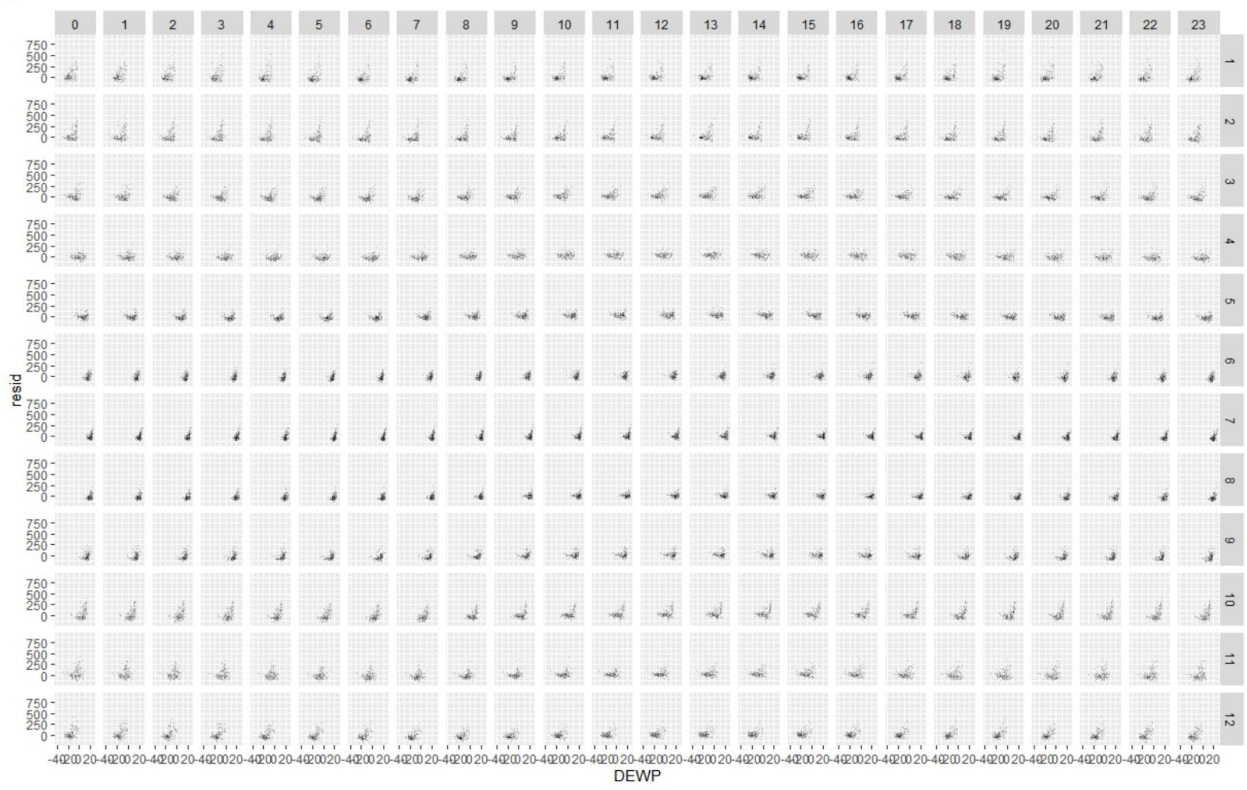
By comparing DEWP and PRES **in conditional to hour**, patterns are similar in each hour, showing a strong negative relationship. So hour is not a significant factor in comparing DEWP and PRES.

4. Can you identify a subset of variables that are effective in terms of predictingpm2.5? Are there differential effects across education, month, hour, and cbwd? Propose a predictive model based on your analysis.

Call:

```
lm(formula = pm2.5 ~ DEWP + TEMP + PRES + month + hour + cbwd +
    Iws + Is + Ir + TEMP * cbwd + TEMP * month + TEMP * hour +
    PRES * cbwd + PRES * month + PRES * hour + DEWP * cbwd +
    DEWP * month + DEWP * hour, data = data)
Residuals:
    Min      1Q  Median      3Q     Max
-180.59  -50.24  -14.80   30.92  907.45
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.330e+03  2.334e+02   5.697 1.23e-08 ***
DEWP         5.668e+00  2.043e-01  27.747  < 2e-16 ***
TEMP        -6.572e+00  2.356e-01 -27.891  < 2e-16 ***
PRES        -1.117e+00  2.286e-01  -4.887 1.03e-06 ***
month       -1.841e+02  2.036e+01  -9.041  < 2e-16 ***
hour         5.002e+01  1.081e+01   4.627 3.72e-06 ***
cbwdNE       8.403e+02  2.583e+02   3.254 0.001139 **
cbwdNW       1.725e+03  1.898e+02   9.089  < 2e-16 ***
cbwdSE      -2.738e+02  2.000e+02  -1.369 0.171098
Iws         -2.149e-01  9.199e-03 -23.365  < 2e-16 ***
Is          -3.293e+00  5.002e-01  -6.583 4.66e-11 ***
Ir          -6.515e+00  2.756e-01 -23.637  < 2e-16 ***
TEMP:cbwdNE -6.320e-02  2.451e-01  -0.258 0.796502
TEMP:cbwdNW  9.802e-01  1.844e-01   5.315 1.07e-07 ***
TEMP:cbwdSE  1.998e+00  1.933e-01  10.340  < 2e-16 ***
TEMP:month  -1.653e-01  1.929e-02  -8.573  < 2e-16 ***
TEMP:hour   -9.079e-03  1.225e-02  -0.741 0.458481
PRES:cbwdNE -8.495e-01  2.525e-01  -3.364 0.000769 ***
PRES:cbwdNW -1.732e+00  1.857e-01  -9.325  < 2e-16 ***
PRES:cbwdSE  2.469e-01  1.956e-01   1.263 0.206747
PRES:month   1.795e-01  1.991e-02   9.016  < 2e-16 ***
PRES:hour   -4.773e-02  1.056e-02  -4.519 6.23e-06 ***
DEWP:cbwdNE  2.459e-01  1.974e-01   1.246 0.212949
DEWP:cbwdNW -1.576e+00  1.516e-01 -10.396  < 2e-16 ***
DEWP:cbwdSE -8.284e-01  1.573e-01  -5.266 1.40e-07 ***
DEWP:month   6.947e-02  1.567e-02   4.434 9.29e-06 ***
DEWP:hour   -4.046e-02  8.653e-03  -4.676 2.93e-06 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 77.95 on 41730 degrees of freedom
Multiple R-squared:  0.2834,  Adjusted R-squared:  0.283
F-statistic: 634.7 on 26 and 41730 DF,  p-value: < 2.2e-16
```

According to the significant values, cbwd, TEMP:cbwd, PRES:cbwd are not very significant, so it is better to drop the cbwd variable.



Residual plots show patterns that indicate that month has a role. Since the residual patterns stay the same for each hour, it is better to drop the interaction between DEWP and hour variable.

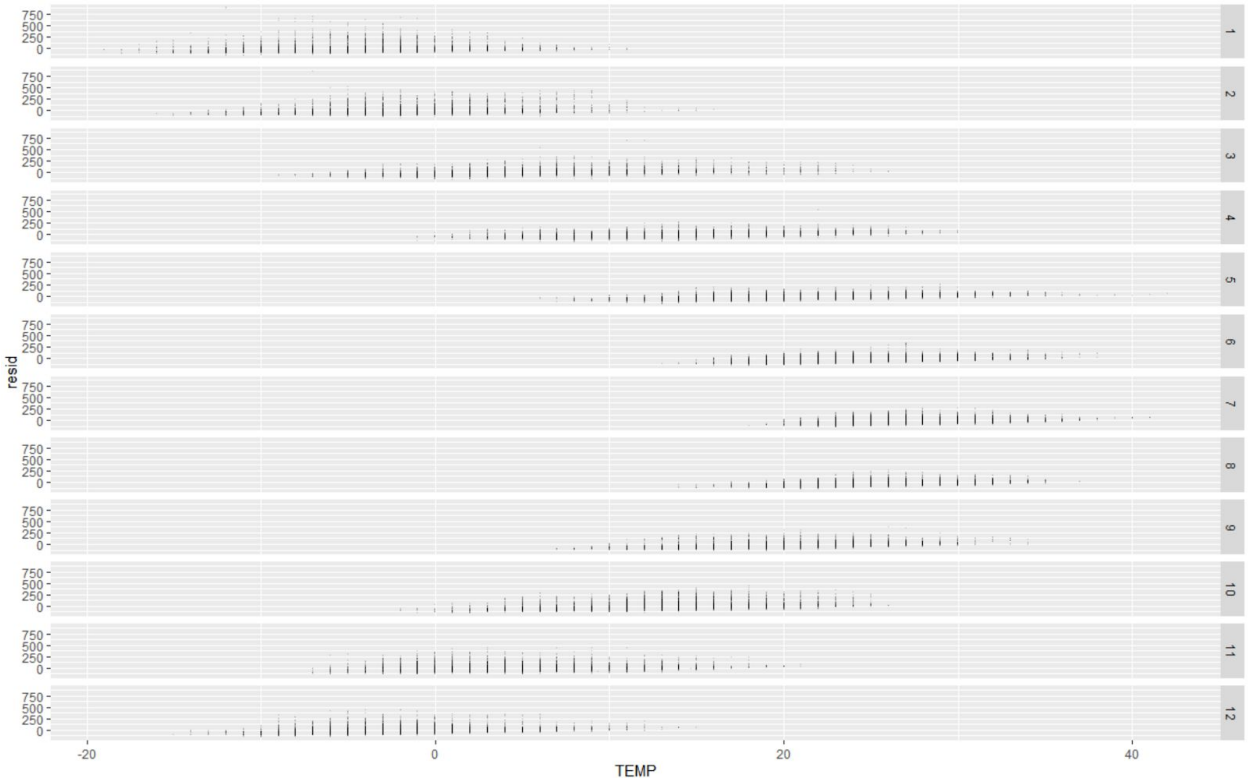
```
Call:
lm(formula = pm2.5 ~ DEWP + TEMP + PRES + month + hour + Iws +
    Is + Ir + TEMP * month + TEMP * hour + PRES * month + PRES *
    hour + DEWP * month, data = data)
Residuals:
    Min      1Q  Median      3Q     Max
-174.07  -52.00  -15.11   32.19  903.25
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.229e+03  1.936e+02  11.517  < 2e-16 ***
DEWP         4.776e+00  1.187e-01  40.223  < 2e-16 ***
TEMP        -5.216e+00  1.687e-01 -30.915  < 2e-16 ***
PRES        -2.024e+00  1.894e-01 -10.689  < 2e-16 ***
month       -1.612e+02  2.052e+01  -7.853 4.17e-15 ***
hour         9.898e+00  1.055e+01   0.938 0.348292
Iws         -2.525e-01  8.382e-03 -30.121  < 2e-16 ***
Is          -2.967e+00  5.046e-01  -5.880 4.12e-09 ***
Ir          -7.403e+00  2.776e-01 -26.669  < 2e-16 ***
TEMP:month  -1.971e-01  1.948e-02 -10.119  < 2e-16 ***
TEMP:hour   -1.489e-02  9.070e-03  -1.641 0.100758
PRES:month   1.568e-01  2.007e-02   7.814 5.69e-15 ***
```

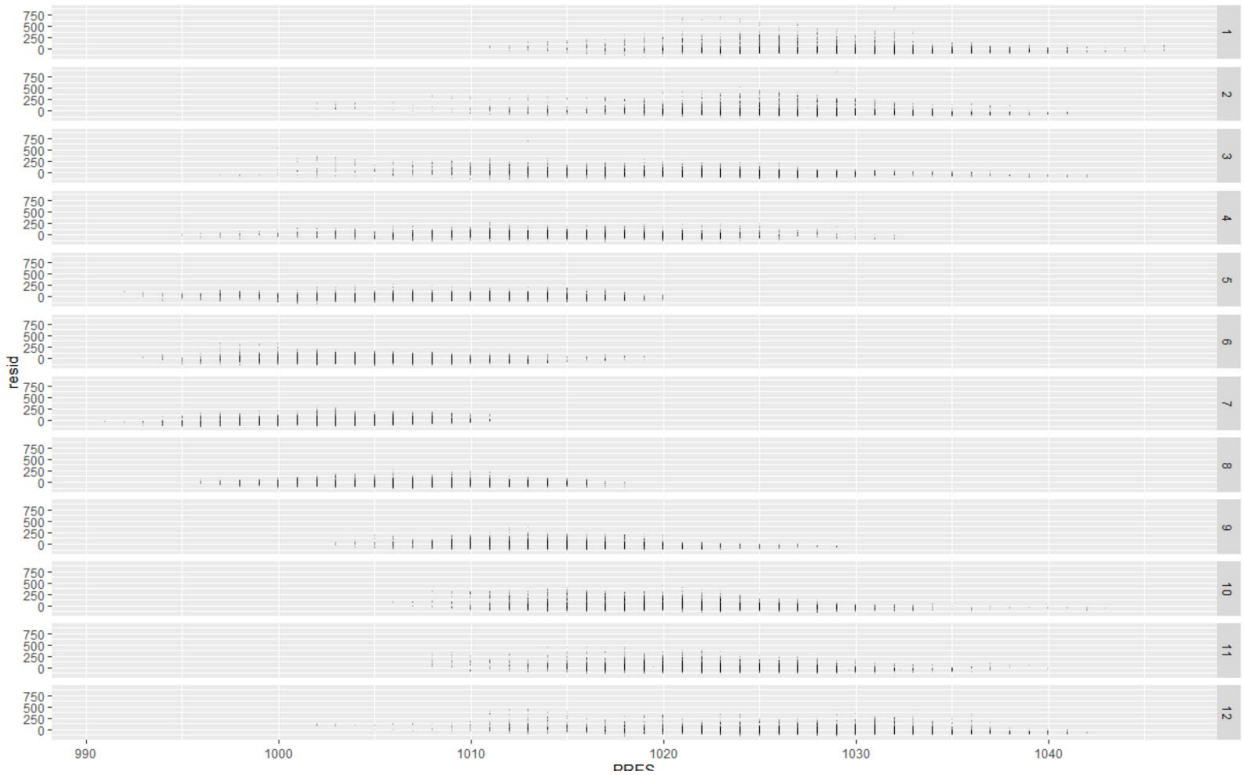
```
PRES:hour   -7.928e-03  1.030e-02  -0.770 0.441415
DEWP:month   5.214e-02  1.576e-02   3.309 0.000937 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.19 on 41743 degrees of freedom
Multiple R-squared:  0.2602,  Adjusted R-squared:   0.26
F-statistic:  1129 on 13 and 41743 DF,  p-value: < 2.2e-16
```

It is better to drop the hour variable since it has a less significant variable

```
Call:
lm(formula = pm2.5 ~ DEWP + TEMP + PRES + month + Iws + Is +
   Ir + TEMP * month + PRES * month + DEWP * month, data = data)

Residuals:
   Min     1Q  Median     3Q    Max
-157.43  -52.16  -14.98   32.32  886.02

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.177e+03  1.592e+02  13.675  < 2e-16 ***
DEWP         4.458e+00  1.193e-01  37.358  < 2e-16 ***
TEMP        -4.888e+00  1.396e-01 -35.026  < 2e-16 ***
PRES        -1.958e+00  1.559e-01 -12.564  < 2e-16 ***
month       -1.585e+02  2.072e+01  -7.651 2.03e-14 ***
Iws         -2.518e-01  8.461e-03 -29.757  < 2e-16 ***
Is          -2.503e+00  5.091e-01  -4.917 8.81e-07 ***
Ir          -7.185e+00  2.802e-01 -25.647  < 2e-16 ***
TEMP:month  -1.847e-01  1.960e-02  -9.423  < 2e-16 ***
PRES:month   1.541e-01  2.025e-02   7.610 2.80e-14 ***
DEWP:month   4.777e-02  1.590e-02   3.004  0.00266 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 79.96 on 41746 degrees of freedom
Multiple R-squared:  0.2457,  Adjusted R-squared:  0.2455
F-statistic:  1360 on 10 and 41746 DF,  p-value: < 2.2e-16
```

Now all the variables remaining are significant





Residual plots show patterns that indicate that month has a role.

So the final model is pm2.5 ~ DEWP + TEMP + PRES + month + Iws + Is + Ir + TEMP * month + PRES * month + DEWP * month, Multiple R-squared: 0.2457, Adjusted R-squared: 0.2455.