Environmental Factors that Influenced Air Quality in New York City in 1973

Manoj Gootam, Deepti Joshi

October 10, 2016

Abstract

Clean air is a basic need and the purpose of assessing air quality is to determine the concentration of air pollutants, impact of quality of air on the environment and the risk to health from air pollutants.An important air pollutant is ozone and there are several factors that result in ozone formation. The aim of the current research was to (a) explore the relationship between ozone levels and environmental factors such as solar radiation, wind speed and temperature (b) examine if factors such as solar radiation, wind speed, temperature and month contribute to an increase in ozone levels, in data collected in peak summer months in 1973, New York city. In order to conduct exploratory and multivariate analyses on 154 observations, we utilized the programming languages, R and Python. Correlation analyses suggest that, consistent with current literature, there is a strong positive correlation between ozone and temperature (r=0.70), strong negative correlation between ozone and wind (r=-0.61) and a moderate positive relationship between ozone and solar radiation (r=0.35). Finally a comparison of two multiple regression models suggested that the cube-root of Ozone levels is linearly related to Temperature, Windspeed and Solar radiation with a total explanatory power of 0.68 (=Adj R-Squared).

Motivation

Clean air is a basic requirement for healthy living and well-being (WHO, 2006). However, depreciating air quality due to an increase in the number of air pollutants continues to threaten our health worldwide. In fact, according to WHO, a significant portion of diseases and premature deaths each year can be attributed to the effects of pollutants in the air. In response to manage rising air pollution, several stakeholders (e.g., World Health Organization, Environmental Protection Agency) have not only identified the key environmental factors that increase air pollution in earth's troposphere, but have also developed guidelines to track and regulate it (Portney, 2016). Notably, a regular analysis of the determinants of air quality is essential since it allows us to develop guidelines for risk assessment, policy development and regulate economic growth in ways that would allow us to control air pollution.

Literature Review

One of the key markers of air quality, and thus a factor of interest for several environmentalists has been the amount of ozone emissions in the earth's ground level atmosphere called the Troposphere (McGarity, 2015). Ozone is a gas that occurs in earth's upper layers of atmosphere as well as in the Troposphere. Ozone is an air pollutant and can be harmful to health and the environment when it is found in high concentrations in the Troposphere. In other words, higher the ozone levels in the Troposphere, lower is the quality of air.Importantly, there are several environmental factors (e.g., solar radiation, wind speed and temperature) that play an important role in the formation of ozone emissions.

Identifying the environmental factors and understanding their relationship with ozone levels is key to undertstand how to regulate and control ozone emissions (West Michigan Clean Air Coalition, 2009). Firstly, solar radiation (i.e., sunlight) has been found to stimulate specific chemical reactions to combine and create the ground-level ozone such that higher the solar radiation, higher are the ozone emissions. Additionally, ground temperatures also influence the formation of ozone such that temperatures higher than 80 degrees Celsius enhance the formation of ozone emissions. Finally, wind speed also contributes to the initiating

of ozone emissions such that wind speeds lower than 10 MPH contribute to the formation of ozone, and higher wind speeds reduce ozone emissions. In conclusion, solar radiation, wind speed, temperature are strong contributors to the development of the ozone emissions. Thus, the research question developed is very specific to identifying the relationship between the ozone levels and the environmental factors that affect ozone formation.

Given the harmful effects of high concentration of ozone emissions in earth's Troposphere it has thus become essential to identify atmospheric conditions that result in high ozone production (Brandt et al., 2016). Specifically, ozone emissions become harmful in atmospheric conditions that increase the formation rate of ozone relative to the dissipation rates. As might be expected, such atmospheric conditions wherein temperatures are high, wind speed is low and solar radiation is high are found typically in the summer months from May to September in the US. Although the reactions leading to ozone formation occur year-round, they only reach harmful levels during the peak summer months. Given that this phenomenon is time-bound, it is relevant for our project to only focus on the peak summer months from May to September. As the harmful effects of high ozone concentration continues to be a major health concern, the insights from current project could be potentially important for policy makers and environmentalists.

Given the significance of environmental factors in contributing to the formation of ozone emissions, it thus is essential to be able to measure, track and regulate these environmental factors as well as ozone emissions (McGarity, 2015). Ozone emissions can be measured by several different instruments that employ different principles and methods (Shao et al., 2015). One such instrument called the 'Ozonesondes,' measures ozone emissions by capturing the light produced in chemical reactions involving ozone. With regards to measuring solar radiation, 'Pyranometers' have been used with great consistency (Huang et al., 2016). Similarly, 'Anemometers' are used to measure wind speed (Baseer et al., 2016). Finally, surface-level temperature is measured with different kinds of thermometers with varied capabilities. In essence, the environmental variables of interest (ozone levels, wind speed, solar radiation and temperature) can be measured with great accuracy and reliability with various instruments employing different techniques. Given the ease at which data can be collected, the research question developed is measurable and answerable

In conclusion, in order to identify and explore the relationship between environmental factors and ozone emissions, the current project utilizes a data set on air quality that consists of 154 observations on ozone emissions and other environemntal factors (e.g., solar radiation, temperature, wind speed, etc) as collected from different locations in the New York city area in the year 1973. The data was originally obtained from the New York State Department of Conservation (ozone data) and the National Weather Service (meteorological data; Chambers et al., 1983).

Exploratory Data Analysis

1. **Descriptive Summary of the data**

Data Summary

The following table is developed based on the information provided in the dataset

Variables	DataType	Units	Time	Location
Ozone	Ratio	Ozone(ppb)	1300 to 1500 hours	Roosevelt Island
Solar Radiation	Ratio	Solar.R(lang)	0800 to 1200 hours	Central Park
Wind Speed	Ratio	Wind(mph)	0700 to 1000 hours	La Guardia Airport
Temperature	Ratio	Temperature(deg F)	N/A	La Guardia Airport
Month	Ordinal	Month(1-12)	N/A	N/A
Day	Ordinal	Day of Month(1-31)	N/A	N/A

Let us start with some descriptive statistics of the dataset

```
## 'data.frame':    153 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 NA 28 23 19 8 NA ...
##  $ Solar.R: int  190 118 149 313 NA NA 299 99 19 194 ...
##  $ Wind   : num  7.4 8 12.6 11.5 14.3 14.9 8.6 13.8 20.1 8.6 ...
##  $ Temp   : int  67 72 74 62 56 66 65 59 61 69 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
##                          Ozone      Solar.R      Wind       Temp
## Mean                 42.099099  184.8018018  9.939640 77.7927928
## Median               31.000000  207.0000000  9.700000 79.0000000
## Mode                 23.000000  238.0000000 11.500000 81.0000000
## Standard Devaition   33.275969   91.1523021  3.557713  9.5299691
## Variance           1107.290090 8308.7421785 12.657324 90.8203112
## Skewness              1.231275   -0.4796906  0.449498 -0.2220609
```

From the table of summary statistics, it can be inferred that solar radiation has the highest variance among all the variables. Additionally, Ozone levels and wind speed are positively skewed while temperature seems to be negatively skewed.

2. **Data Cleaning**

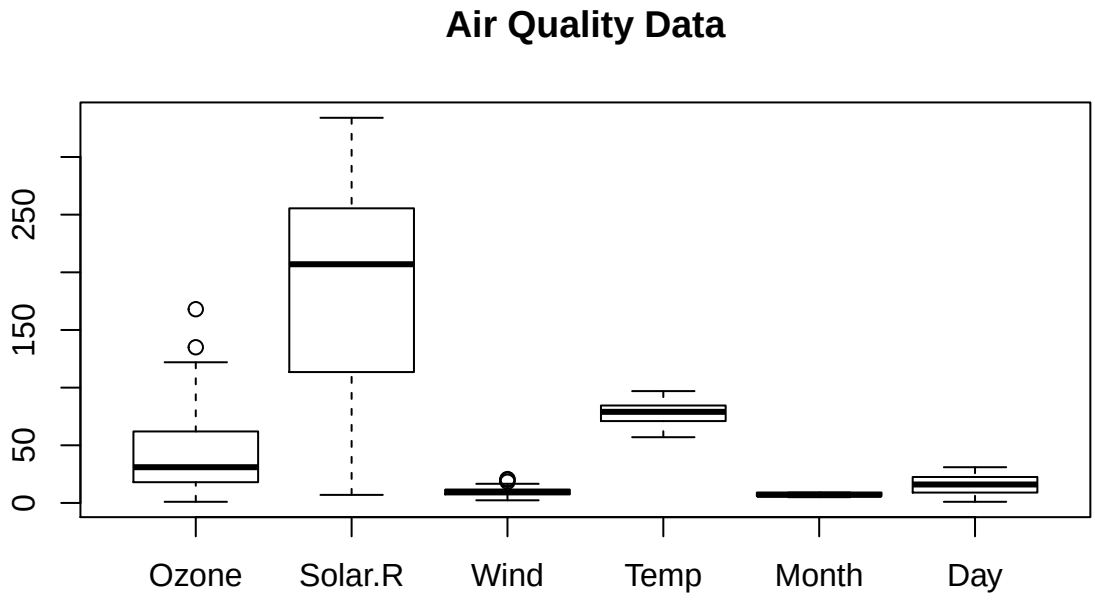
While there are some observations with extreme values, they do not necessarily qualify as outliers. Throughout the exercise, it has been assumed that outliers are observations which are 4 standard deviations away fronm the mean. In this particular dataset, no observation is beyond 4 standard deviations. So there are no outliers. However, there are about 42 observations with the values as NAs which were removed in the following section before going through further analysis.

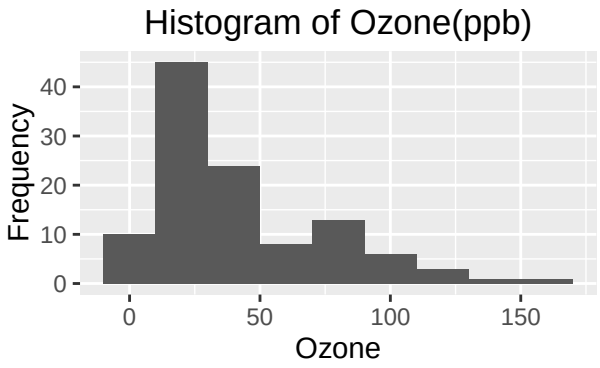
```
## Data Cleaning: There are some observations with NAs in it, which have to be removed.
indices.NA = apply(airquality,1,anyNA)
airquality.NA = airquality[indices.NA,]
airquality = airquality[!indices.NA,]

if(!anyNA(airquality)){
  cat(str(airquality))
}
```

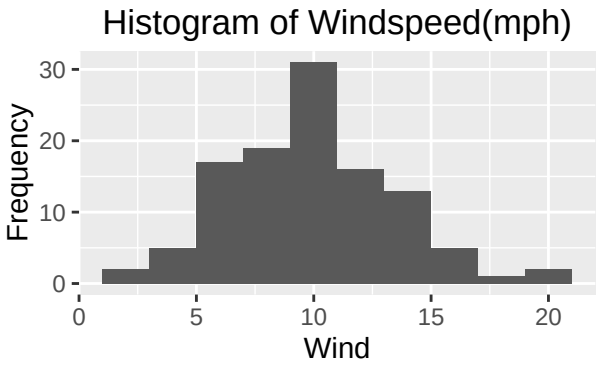
```
## 'data.frame':    111 obs. of  6 variables:
##  $ Ozone  : int  41 36 12 18 23 19 8 16 11 14 ...
##  $ Solar.R: int  190 118 149 313 299 99 19 256 290 274 ...
##  $ Wind   : num  7.4 8 12.6 11.5 8.6 13.8 20.1 9.7 9.2 10.9 ...
##  $ Temp   : int  67 72 74 62 65 59 61 69 66 68 ...
##  $ Month  : int  5 5 5 5 5 5 5 5 5 5 ...
##  $ Day    : int  1 2 3 4 7 8 9 12 13 14 ...
```

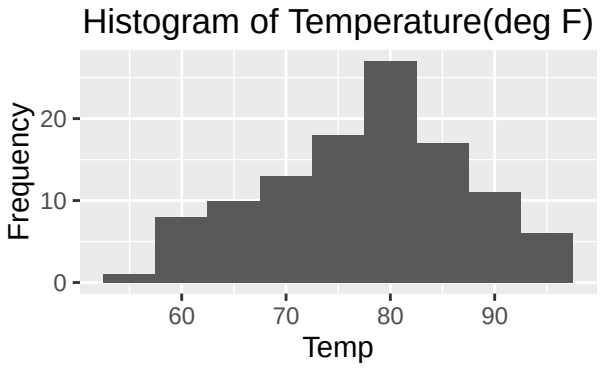
3. **Distributions of all the variables**

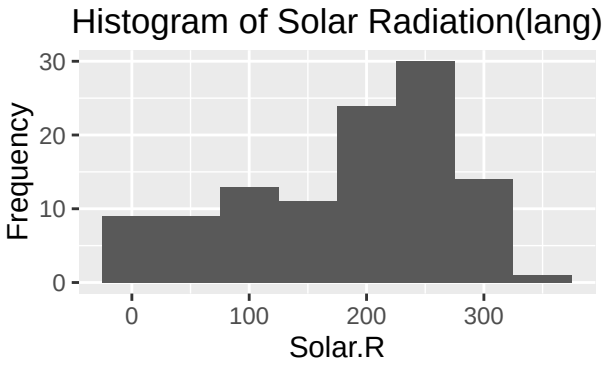
In order to understand the distribution of each of the variables, the following visualizations were conducted. The following boxplots and series of histograms, suggest there is great variability in solar radiation, followed by Ozone levels. Additionally, Ozone is strongly positively skewed while wind speed is moderately positvely skewed. Finally, solar radiation is negatively but moderately skewed as seen in the histograms and boxplot.

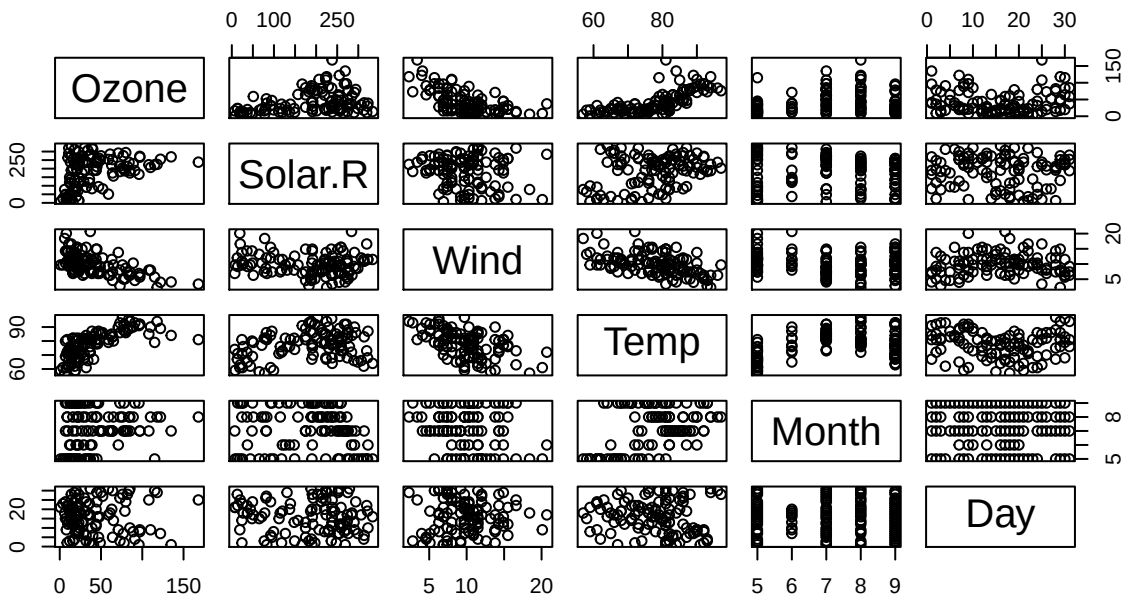

Air Quality Data
0
50
150
250
Ozone
Solar.R
Wind
Temp
Month
Day


Histogram of Ozone(ppb)
Frequency
0
10
20
30
40
0
50
100
150
Ozone

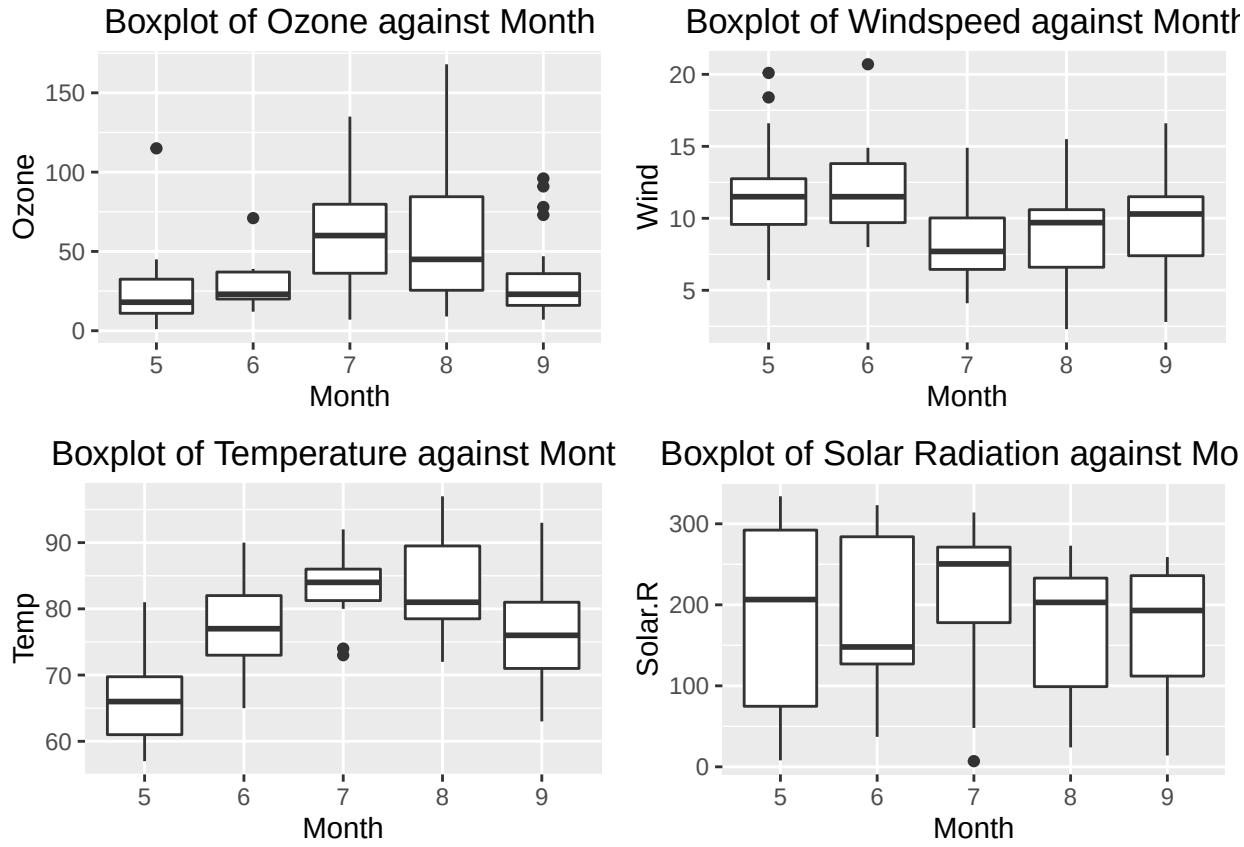

Histogram of Windspeed(mph)
Frequency
0
10
20
30
0
5
10
15
20
Wind


Histogram of Temperature(deg F)
Frequency
0
10
20
60
70
80
90
Temp

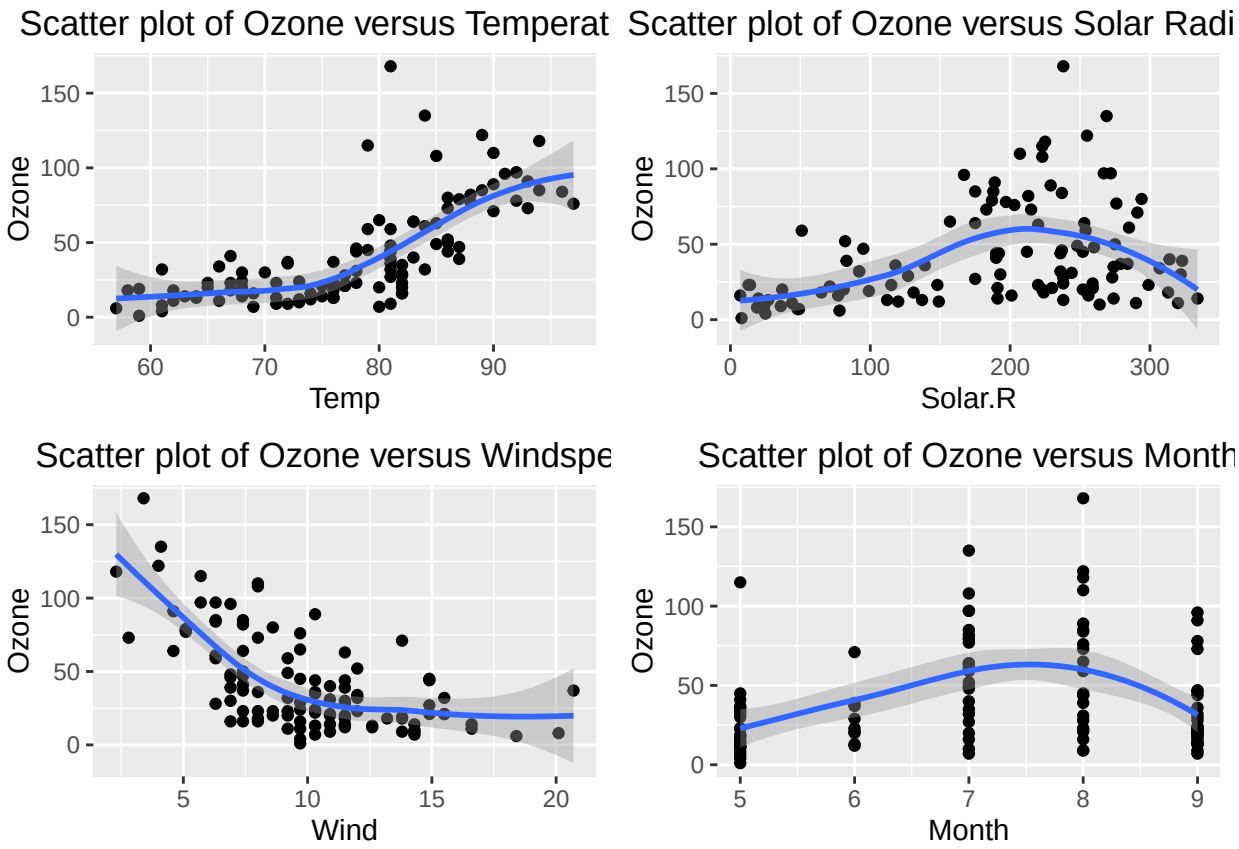

Histogram of Solar Radiation(lang)
Frequency
0
10
20
30
0
100
200
300
Solar.R



The pairwise scatter plots suggest that ozone, wind speed, temperature and solar radiation have a specific pattern against Month. Moreover, ozone and temperature seem to be positively correlated while ozone is negatively correlated with temperature. Surpsrisingly, no clear pattern can be deduced between ozone and solar radiation.



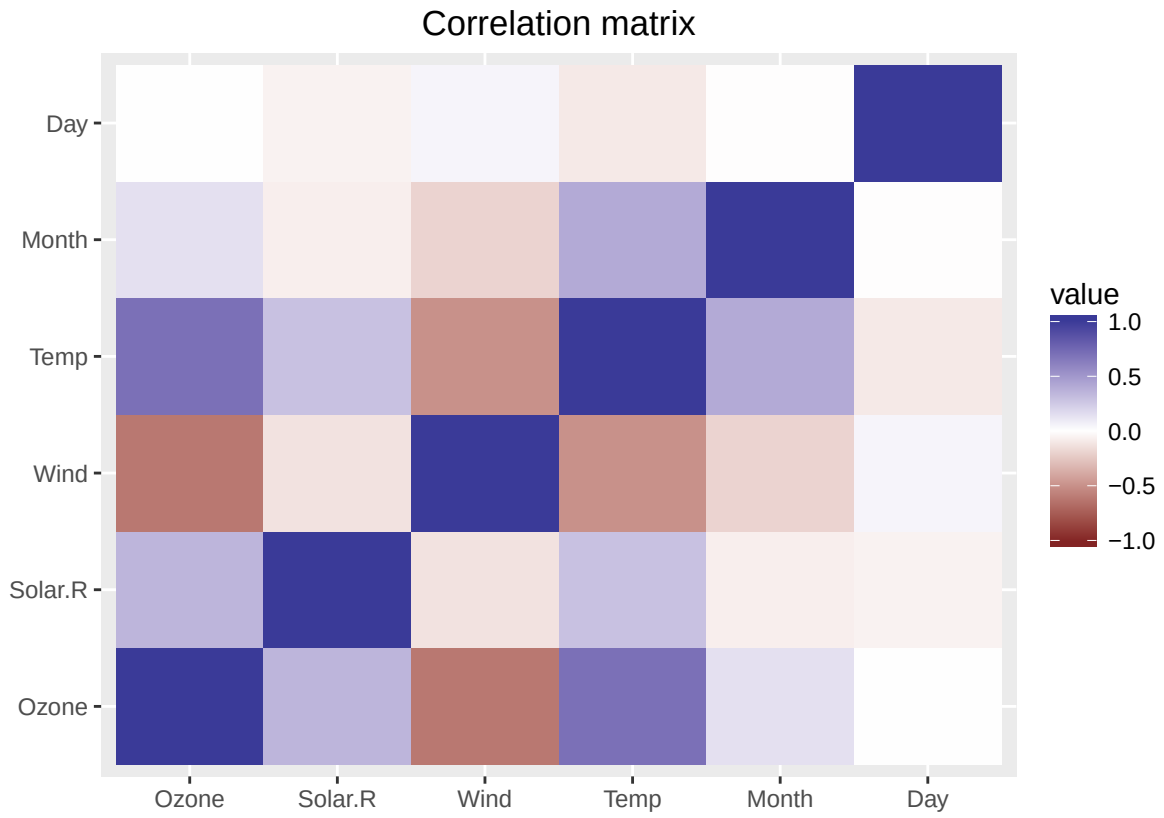
From the above boxplots it can be clearly observed that the Ozone levels increase up until July and starts to fall down in the month of august and september. Temperature also exhibits similar behaviour. However, Wind speed reduces through the month of july and starts to increase through september. Due to the variability in Solar Radiation values, no clear pattern is observed.



From the above scatter plots it can be inferred that Ozone tends to have a strong positive correlation with Temperature, strong negative correlation with Windspeed and a moderate positive correlation with Solar Radiation

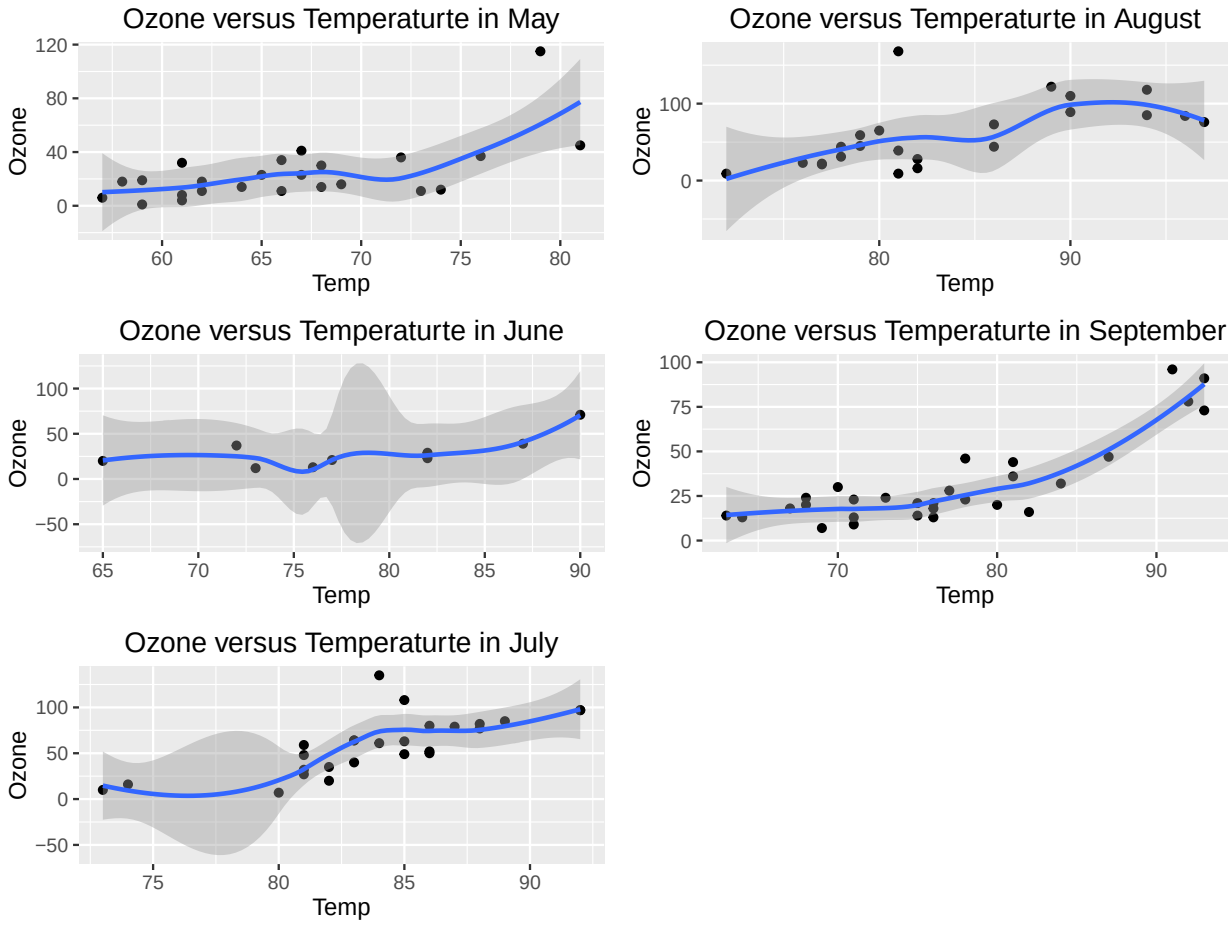
3.1 **Correlation analysis**

```
##                Ozone     Solar.R        Wind       Temp        Month
## Ozone    1.000000000  0.34834169 -0.61249658  0.6985414  0.142885168
## Solar.R  0.348341693  1.00000000 -0.12718345  0.2940876 -0.074066683
## Wind    -0.612496576 -0.12718345  1.00000000 -0.4971897 -0.194495804
## Temp     0.698541410  0.29408764 -0.49718972  1.0000000  0.403971709
## Month    0.142885168 -0.07406668 -0.19449580  0.4039717  1.000000000
## Day     -0.005189769 -0.05775380  0.04987102 -0.0965458 -0.009001079
##                 Day
## Ozone   -0.005189769
## Solar.R -0.057753801
## Wind     0.049871017
## Temp    -0.096545800
## Month   -0.009001079
## Day      1.000000000
```

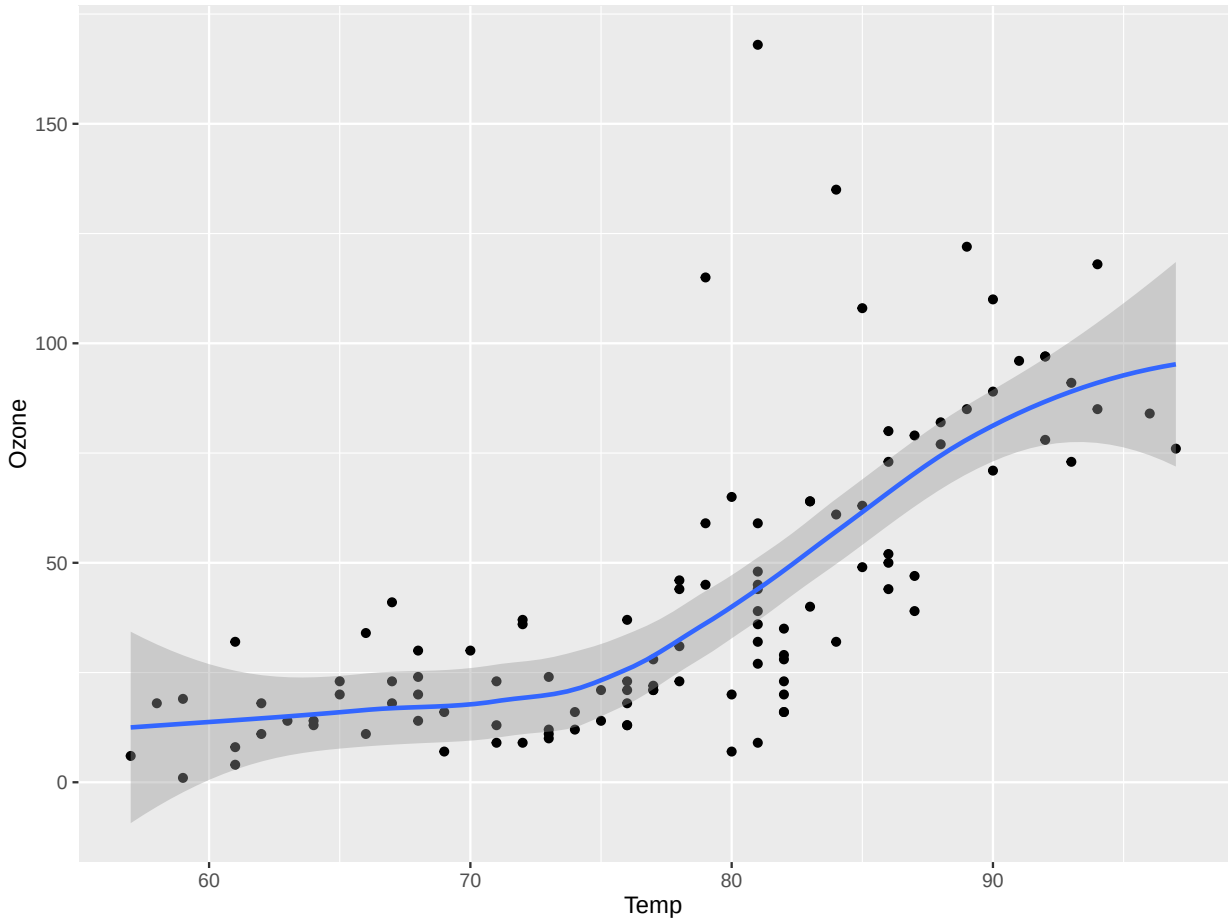


As expected from current literature, there is a strong positive correlation between ozone and temperature (r=0.70), strong negative correlation between ozone and wind (r=-0.61), moderate positive relationship between ozone and solar radiation (r=0.35) and weak correlation between ozone and month (r=0.14)

```
P5 = ggplot(airquality[which(airquality$Month==5),], aes(Temp, Ozone)) + geom_point()+ geom_smooth()+xla
P6 = ggplot(airquality[which(airquality$Month==6),], aes(Temp, Ozone)) + geom_point()+ geom_smooth()+xla
P7 = ggplot(airquality[which(airquality$Month==7),], aes(Temp, Ozone)) + geom_point()+ geom_smooth()+xla
P8 = ggplot(airquality[which(airquality$Month==8),], aes(Temp, Ozone)) + geom_point()+ geom_smooth()+xla
P9 = ggplot(airquality[which(airquality$Month==9),], aes(Temp, Ozone)) + geom_point()+ geom_smooth()+xla
multiplot(P5, P6, P7, P8, P9, cols=2)
```

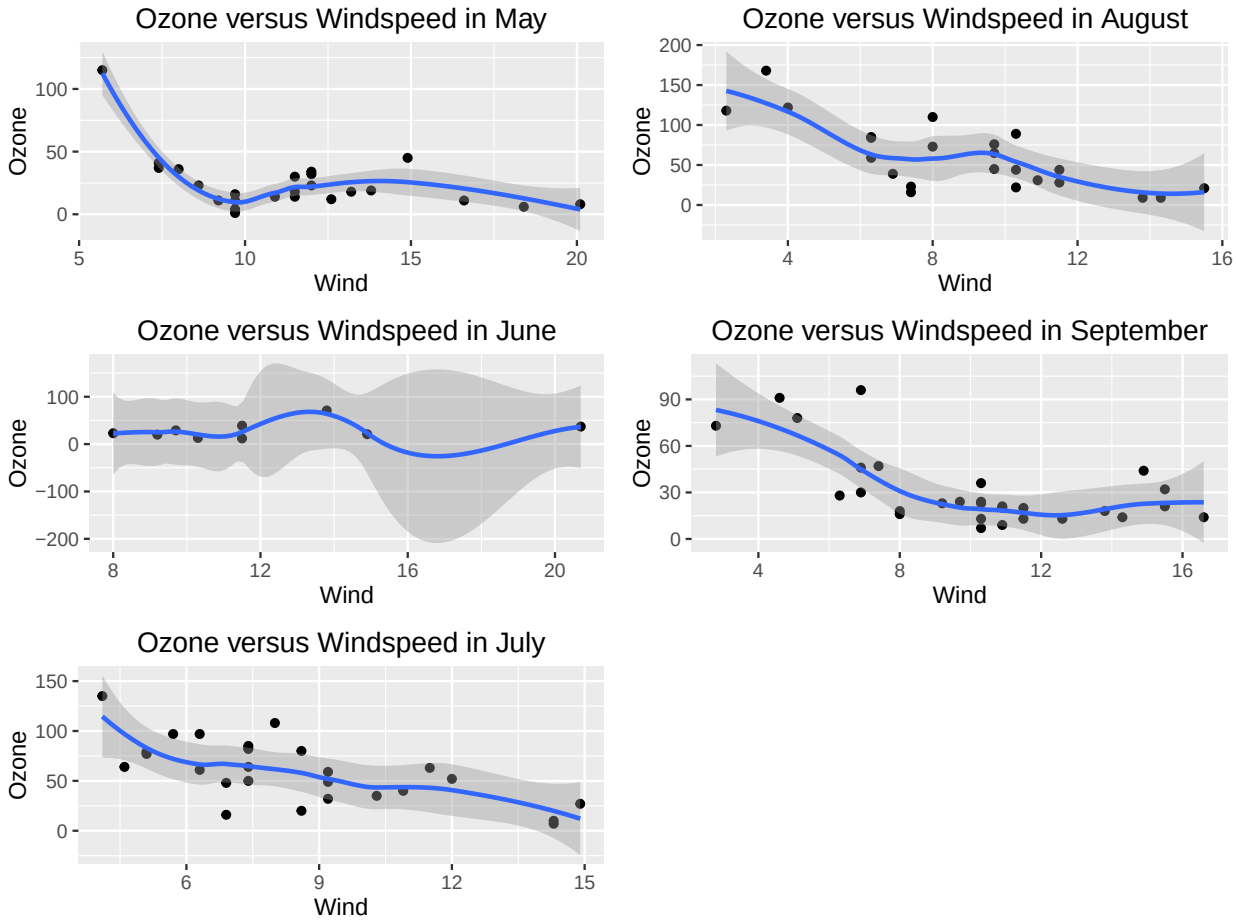



```
ggplot(airquality, aes(Temp, Ozone)) + geom_point()+ geom_smooth()
```

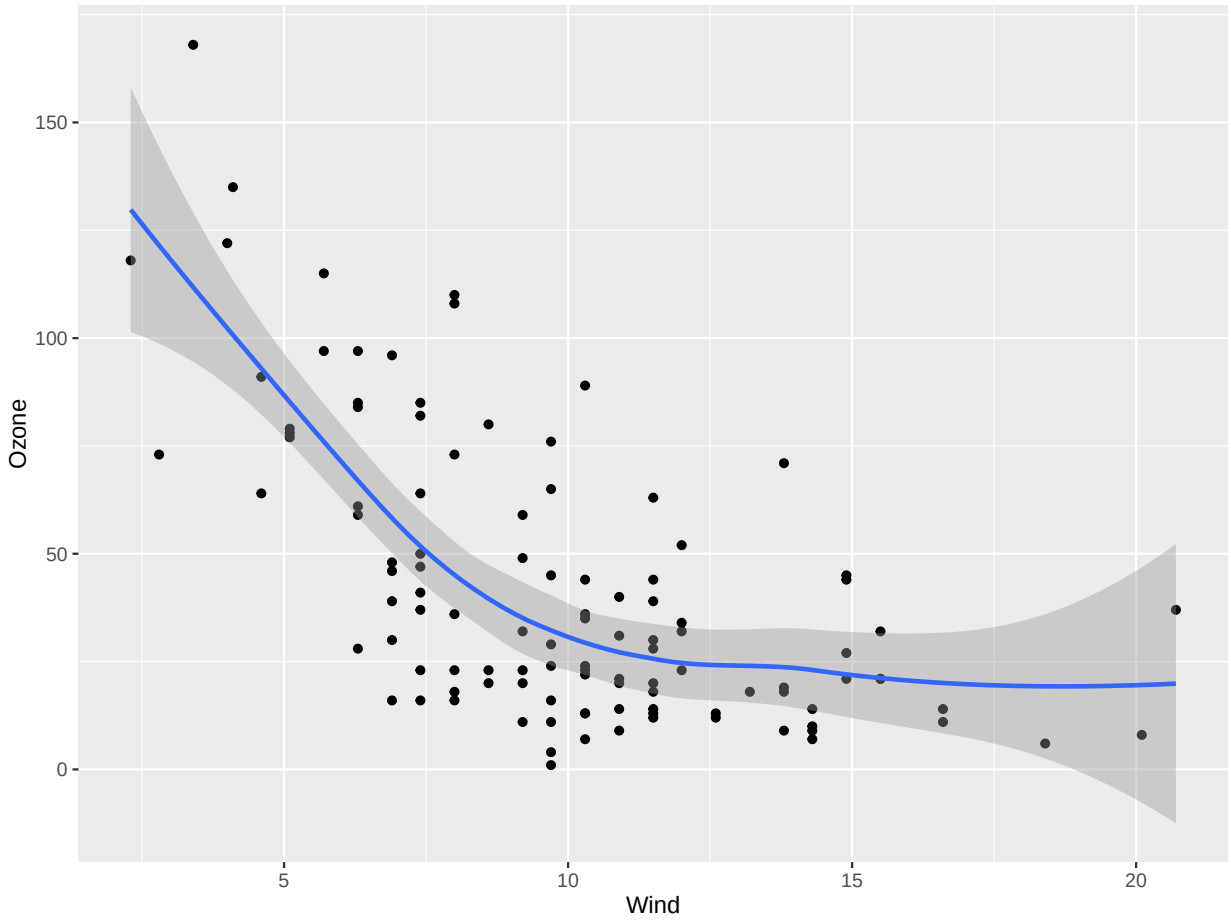


From the above scatter plots it can be inferred that Ozone levels increase with increase in temperature establishing strong positive correlation consitent with the literature.

```
P5 = ggplot(airquality[which(airquality$Month==5),], aes(Wind, Ozone)) + geom_point()+ geom_smooth()+xla
P6 = ggplot(airquality[which(airquality$Month==6),], aes(Wind, Ozone)) + geom_point()+ geom_smooth()+xla
P7 = ggplot(airquality[which(airquality$Month==7),], aes(Wind, Ozone)) + geom_point()+ geom_smooth()+xla
P8 = ggplot(airquality[which(airquality$Month==8),], aes(Wind, Ozone)) + geom_point()+ geom_smooth()+xla
P9 = ggplot(airquality[which(airquality$Month==9),], aes(Wind, Ozone)) + geom_point()+ geom_smooth()+xla
multiplot(P5, P6, P7, P8, P9, cols=2)
```




```
ggplot(airquality, aes(Wind, Ozone)) + geom_point()+ geom_smooth()
```



From the above scatter plots it can be inferred that Ozone levels decrease with increase in Windspeed establishing strong negative correlation consitent with the literature.

3.2 **Linear Regression**

```
##
## Call:
## lm(formula = Ozone ~ ., data = airquality)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.014 -12.284  -3.302   8.454  95.348
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -64.11632   23.48249  -2.730  0.00742 **
## Solar.R       0.05027    0.02342   2.147  0.03411 *
## Wind         -3.31844    0.64451  -5.149 1.23e-06 ***
## Temp          1.89579    0.27389   6.922 3.66e-10 ***
## Month        -3.03996    1.51346  -2.009  0.04714 *
## Day           0.27388    0.22967   1.192  0.23576
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.86 on 105 degrees of freedom
```

```
## Multiple R-squared:  0.6249, Adjusted R-squared:  0.6071
## F-statistic: 34.99 on 5 and 105 DF,  p-value: < 2.2e-16
```

The first multiple regression model produced an adjusted R-squared value of 0.62, implying that this set of independent variables (Solar radiation, wind speed, temperature and month) explain about 62 % of the variance in ozone levels.The R-squared value of 0.62 suggests that the model is a good fit for the data.Additionally, the residual standard error is relatively small as it only amounts to about 30% of the variance in the ozone levels. As seen from the table, temperature has significant positive regression coefficients indicating that the ozone formation is expected to rise by 1.9ppb with 1 unit increase in temperature.In other words, temperature is correlated with ozone levels after controlling for all other variables.Similarly, wind also has a significant but negative regression coefficient indicating that the ozone formation is expected to fall by 3.32ppb with 1 unit increase in wind speed.In other words, wind speed is correlated with ozone levels after controlling for all other variables. Similarly, solar radiation also has a significant positive regression coefficent indicating that ozone formation is expected to rise by 0.05ppb with a 1 unit increase in solar radiation. In other words, solar radiation is correlated with ozone levels after controlling for all other variables. Finally, month also has a significant negative regression coefficent indicating that ozone formation is expected to fall by 3.05 ppb with every month. In other words, month is correlated with ozone levels after controlling for all other variables. The following is the regression equation for the current model: Ozone = b0 + b1 $Temp$ + $b2$Wind + b3$Solar.R$ + $b4$Month + b5*Day

```
airquality$Ozone = airquality$Ozone^(1/3)
attach(airquality)
```

```
## The following objects are masked from airquality (pos = 4):
##
##     Day, Month, Ozone, Solar.R, Temp, Wind
```

```
print(cor(airquality))
```

```
##              Ozone     Solar.R        Wind       Temp        Month
## Ozone    1.0000000  0.42201296 -0.59864094  0.7531038  0.175495031
## Solar.R  0.4220130  1.00000000 -0.12718345  0.2940876 -0.074066683
## Wind    -0.5986409 -0.12718345  1.00000000 -0.4971897 -0.194495804
## Temp     0.7531038  0.29408764 -0.49718972  1.0000000  0.403971709
## Month    0.1754950 -0.07406668 -0.19449580  0.4039717  1.000000000
## Day     -0.0249254 -0.05775380  0.04987102 -0.0965458 -0.009001079
##                  Day
## Ozone   -0.024925399
## Solar.R -0.057753801
## Wind     0.049871017
## Temp    -0.096545800
## Month   -0.009001079
## Day      1.000000000
```

```
print(summary(lm(Ozone~.,data = airquality)))
```

```
##
## Call:
## lm(formula = Ozone ~ ., data = airquality)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.04894 -0.34853  0.02008  0.30779  1.48816
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.2976191  0.5688921  -0.523 0.601968
## Solar.R      0.0020048  0.0005673   3.534 0.000611 ***
## Wind        -0.0756482  0.0156140  -4.845 4.39e-06 ***
## Temp         0.0552204  0.0066352   8.322 3.42e-13 ***
## Month       -0.0642541  0.0366653  -1.752 0.082616 .
## Day          0.0059419  0.0055641   1.068 0.288009
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5053 on 105 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6779
## F-statistic: 47.31 on 5 and 105 DF,  p-value: < 2.2e-16
```

The second multiple regression model produced an adjusted R-squared value of 0.68, implying that the set of independent variables (Solar radiation, wind speed, temperature) explain about 68 % of the variance in ozone levels.The R-squared value of 0.68 suggests that the model is a good fit for the data.Additionally, the residual standard error is relatively small as it only amounts to about 0.5. As seen from the table, temperature has significant positive regression coefficients indicating that the ozone formation is expected to rise by 0.06ppb with 1 unit increase in temperature.In other words, temperature is correlated with ozone levels after controlling for all other variables.Similarly, wind speed also has a significant but negative regression coefficient indicating that the ozone formation is expected to fall by 0.08ppb with 1 unit increase in wind speed.In other words, wind speed is correlated with ozone levels after controlling for all other variables. Similarly, solar radiation also has a significant positive regression coefficent indicating that ozone formation is expected to rise by 0.002ppb with a 1 unit increase in solar radiation. In other words, solar radiation is correlated with ozone levels after controlling for all other variables. The following is the regression equation for the current model: Ozoneˆ1/3 = b0 + b1*Temp* + *b2*Wind speed + b3*Solar.R* + *b4*Month + b5*Day

Conclusion

Comparing the two regression models it seems as though the Model2 has lesser residual standard error compared to Model 1 indicating that the lack of fit is high for model 1. While model 2 explains 68% of the ozone levels, model 1 explains only 60% of the Ozone levels which implies that model 2 has a better Goodness of fit.Additionally, model2 has statistically significant regression coefficients with high confidence percentage when compared to model 1. For all the above reasons, Model2 would be a better regression model. Finally, in both the models, condition number is large implying the presence of multicollinearity, which has to be addressed in the future. In conclusion, solar radiation, wind speed and temperature were significant predictors of ozone in New york in 1973.

References

1. Baseer, M. A., Meyer, J. P., Rehman, S., Alam, M. M., Al-Hadhrami, L. M., & Lashin, A. (2016). Performance evaluation of cup-anemometers and wind speed characteristics analysis. Renewable Energy, 86, 733-744.Anemometers. Review of Scientific Instruments, 38(5), 677-681.
2. Brandt, R. E., Schwab, J. J., Casson, P. W., Roychowdhury, U. K., Wolfe, D., Demerjian, K. L., ... & Felton, H. D. (2016). Atmospheric Chemistry Measurements at Whiteface Mountain, NY: Ozone and Reactive Trace Gases. Aerosol Air Qual. Res, 16, 873-884.

3. Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983) Graphical Methods for Data Analysis. Belmont, CA: Wadsworth.
4. Huang, G., Li, X., Huang, C., Liu, S., Ma, Y., & Chen, H. (2016). Representativeness errors of point-scale ground-based solar radiation measurements in the validation of remote sensing products. Remote Sensing of Environment, 181, 198-206.
5. Lin, X., Pielke Sr, R. A., Mahmood, R., Fiebrich, C. A., & Aiken, R. (2016). Observational evidence of temperature trends at two levels in the surface layer. Atmospheric Chemistry and Physics, 16(2), 827-841.
6. McGarity, T. O. (2015). Science and Policy in Setting National Ambient Air Quality Standards: Resolving the Ozone Enigma. Texas Law Review, 93.
7. World Health Organization. (2006). Air quality guidelines: global update 2005: particulate matter, ozone, nitrogen dioxide, and sulfur dioxide. World Health Organization.
8. Factors that Contribute to the Formation of Ozone and Particulate Matter. West Michigan Clean Air Coalition (2009, January 22). Retrieved October 08, 2016, from http://www.wmcac.org/airquality/factors.html
9. Anemometer. (n.d.). Retrieved October 08, 2016,from https://en.wikipedia.org/wiki/Anemometer
10. Venkanna, R., Nikhil, G. N., Sinha, P. R., Rao, T. S., & Swamy, Y. V. (2016). Significance of volatile organic compounds and oxides of nitrogen on surface ozone formation at semi-arid tropical urban site, Hyderabad, India. Air Quality, Atmosphere & Health, 9(4), 379-390.