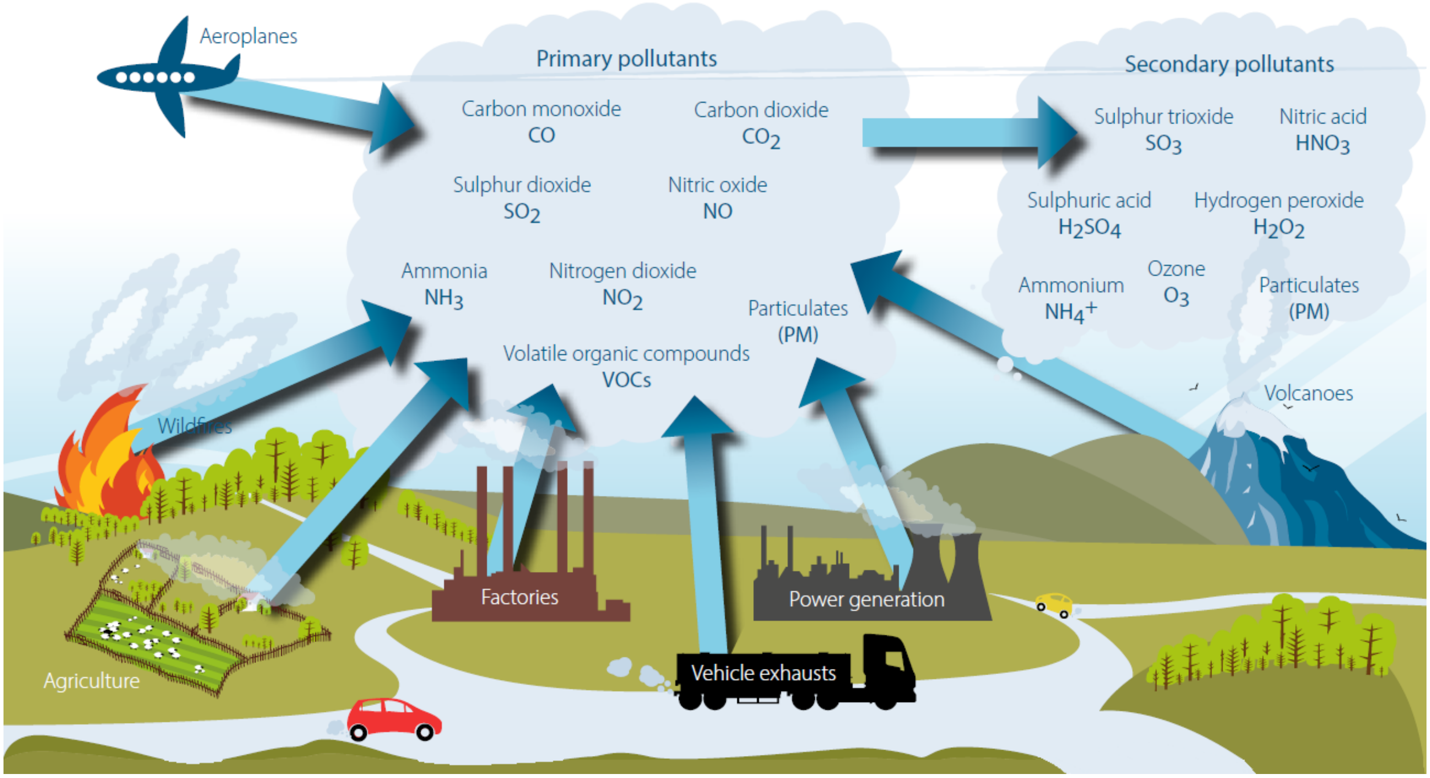
DATA 606 Final Project

Samantha Deokinanan

5th May 2019

Introduction

Clean air is vital as it provides oxygen and other gases that sustain the delicate balance of life on Earth. However, the quality of the air can be affected by air pollution. Air pollution occurs when certain gases and particles build up in the atmosphere to such levels that they can cause disease, death to humans, damage to other living organisms such as food crops, or damage to the natural or man-made environment. These substances, known as pollutants, can be solid particles, liquid droplets, or gases, and are classified as primary or secondary pollutants. The primary pollutant tends to come from man-made sources, including the burning of fossil fuels such as coal, oil, petrol or diesel, but can also come from natural sources such as volcanic eruptions and forest fires. Unlike primary pollutants, secondary pollutants are not emitted directly. Rather, they form in the air when primary pollutants as a result of chemical reactions. The federal Clean Air Act authorized the Environmental Protection Agency (EPA) to set National Ambient Air Quality Standards (NAAQS) for pollutants that threaten human health and public welfare throughout the country (*Clean Air Act, EPA*). EPA established NAAQS for six most common pollutants called "criteria" air pollutants: ground-level ozone (O_3), fine particulate matter ($PM_2.5$ and PM_10), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2), and lead (Pb), among which ground level O_3, $PM_2.5$ and NO_2 (main component of NO_x) are the most widespread health threats.



SEPA: The chemistry of air pollution

Primary Pollutants

The primary pollutant, carbon monoxide is one of the most harmful air pollutants. It is majorly produced from motor vehicle exhaust, along with other primary pollutants such as nitrogen oxides (eg nitrogen dioxide, NO_2). Particulate matter is a complex mixture of organic and inorganic substances, present in the atmosphere as both liquids and solids. Coarse particulates can be regarded as those with a diameter greater than 2.5 micrometers (μm), usually contain earth crustal materials and dust from road vehicles and industries, and fine particles less than 2.5 μm, contains aerosols, combustion particles, and re-condensed organic and metallic vapors. Another primary pollutant is sulfur dioxide (SO_2), released from power stations and industrial plants.

Air pollution has a serious toxicological impact on human health and the environment. CO is poisonous when inhaled because it combines faster with hemoglobin, the oxygen-carrying substance in red blood cells, than oxygen. As a result, the lack of oxygen causes cells and tissues to die. Similarly, in significant concentrations, nitrogen dioxide is highly toxic, causing serious lung damage with a delayed effect. It also plays a major role in the atmospheric reactions that produce ground-level O_3 or smog. Whereas fine particles of less than 10 μm in diameter can penetrate deep into the lung and cause more damage, and SO_2 pollution is known to cause heart disease and bronchitis. Pollutants have even more adverse effect when in moderate concentrations as it can lead to a fall in lung function in asthmatics. SO_2 pollution is considered more harmful when particulate and other pollution concentrations are high. This is known as the “cocktail effect.” Increasing concentration of these gases in the atmosphere also causes global warming and climate change as N_2O has 310 times more global warming potentiality than CO_2 (*Sen et al, 2017*).

Secondary Pollutants

Ground level O_3 is a prominent example of a secondary pollutant, formed by the action of sunlight on volatile organic compounds such as Benzene (C_6H_6) in the presence of NO_2. There are no direct man-made emissions of O_3 to the atmosphere. Ozone can cause irritation to the respiratory tract and eyes, causing chest tightness, coughing and wheezing, especially amongst those with respiratory and heart problems. Ground level O_3 can also have detrimental effects on plants and ecosystems, including damage to plants, reductions of crop yield, and an increase of vegetation vulnerability to disease (*Criteria Air Pollution, EPA*).

Research Objective

As air pollution is a complex mixture of toxic components with considerable impact on humans, many experts claim that forecasting air pollution concentration is a priority for improving life quality (*De Vito et al, 2008; Peng, 2015; Sen et al, 2017*). The goals of this project are to investigate the following:

- What are the predictors that affect the level of specific pollutants in the air? How impactful is their presence?
- With the significant predictor(s) that affects the levels of pollutants in the air, how do they change based on the season? Is a season more prone to more emission of one or more of a specific air pollutant than another?
- When non-metallic hydrocarbon are combusted, they produce CO. With the limited data on NMHC concentration (90% missing values), is NMHC still a contributor in predicting the level of CO in the air given this data?

Data

The data was collected by Saverio De Vito (saverio.devito '@' enea.it) from ENEA - National Agency for New Technologies, Energy and Sustainable Economic Development. It was then submitted to the University of California Irvine, School of Information and Computer Science, Machine Learning Repository. It contains 9358 instances of hourly averaged responses from an array of 5 metal oxide chemical sensors embedded in an Air Quality Chemical Multi-sensor Device. The device was located on the field in a significantly polluted area, at road level, within an Italian city, thus making this an observational study. Data were recorded from March 2004 to February 2005 (one year) representing the longest freely available recordings of on-field deployed air quality chemical sensor devices responses.

The response variables are the 10 hourly averaged responses from an Air Quality Chemical Multisensor Device, in addition to Temperature, Relative Humidity, and Absolute Humidity records. All of which are quantitative. The independent variables are the data and time the responses were recorded. While `Time` is quantitative, `Date` was in both quantitative and qualitative formats for different analyses. `Date` was tidied and transformed into qualitative variables `Season` , and `MonthName` . The variables of the **original** data set are:

1. Date (DD/MM/YYYY)
2. Time (HH.MM.SS)
3. True hourly averaged concentration CO in $microg/m^3$ (reference analyzer)
4. PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)
5. True hourly averaged overall Non-Metallic HydroCarbons concentration in $microg/m^3$ (reference analyzer)

6. True hourly averaged Benzene concentration in $microg/m^3$ (reference analyzer)
7. PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)
8. True hourly averaged NOx concentration in ppb (reference analyzer)
9. PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)
10. True hourly averaged NO_2 concentration in $microg/m^3$ (reference analyzer)
11. PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO_2 targeted)
12. PT08.S5 (indium oxide) hourly averaged sensor response (nominally O_3 targeted)
13. Temperature in $^\circ C$
14. Relative Humidity (%)
15. AH Absolute Humidity

Building a predictive quantitative tool that can easily be used and provide valuable information on individual pollutants and one-hour average concentration will make it more flexible to future changes. Such results may also be beneficial to other researchers, environmentalist, enthusiasts, and meteorologists. While these results cannot be generalized to all populations, with collected air quality data from any location, the methods can be replicated to determine the predictors that affect the level of specific pollutants in the air those locations. Moreover, this data cannot be used to establish causal links between the variables of interest. Poor air quality is associated with serious toxicological impact on human health and the environment, however, studies have shown that there is no overall reduction in mortality with improved air quality, thus causality is not supported.

Exploratory Data Analysis

The data set is retrieved in its raw form, and have the independence of observation since they were collected daily. Therefore some data tidying and transformation are conducted, in addition to exploratory data analysis.

Data Tidy & Transformation

```
# The required R packages
library(tidyverse)
library(lubridate)
library(MASS)
library(caret)
library(olsrr)
library(psych)
library(rcompanion)
library(mctest)

# Load the .csv file from local machine
AirQualityUCI <- read_csv("AirQualityUCI.csv", col_types = cols(AH = col_number(), `C6H6(GT)`
 = col_number(), `CO(GT)` = col_number(), Date = col_date(format = "%m/%d/%Y"), `NMHC(GT)` = c
ol_number(), `NO2(GT)` = col_number(), `NOx(GT)` = col_number(), `PT08.S1(CO)` = col_number(),
`PT08.S2(NMHC)` = col_number(), `PT08.S3(NOx)` = col_number(), `PT08.S4(NO2)` = col_number(),
 `PT08.S5(O3)` = col_number(), RH = col_number(), T = col_number(), Time = col_time(format =
"%H:%M:%S")))

# Identifying seasons and month names, and split date into year, month, and day
## Month Name
AirQualityUCI$MonthName <- month(ymd(AirQualityUCI$Date), label = TRUE, abbr = FALSE)

## Identify the season
Season <- function(Date) {
    Winter <- as.Date("2003-12-20", format = "%Y-%m-%d")
    Spring <- as.Date("2004-3-20",  format = "%Y-%m-%d")
    Summer <- as.Date("2004-6-20",  format = "%Y-%m-%d")
    Fall <- as.Date("2004-9-20",  format = "%Y-%m-%d")
    Winter2 <- as.Date("2004-12-20", format = "%Y-%m-%d")
    ifelse (Date >= Winter & Date < Spring, "Winter 04",
      ifelse (Date >= Spring & Date < Summer, "Spring 04",
        ifelse (Date >= Summer & Date < Fall, "Summer 04",
          ifelse (Date >= Fall & Date < Winter2, "Fall 04", "Winter 05"))))
}

AirQualityUCI$Season <- Season(AirQualityUCI$Date)

## Split date into year, month and day
AirQualityUCI <- AirQualityUCI %>%
  separate(Date, sep="-", into = c("Year", "Month", "Day"))

# Missing values (indicated by -200) reassigned to NA
AirQualityUCI[AirQualityUCI == -200] <- NA

str(AirQualityUCI)
```

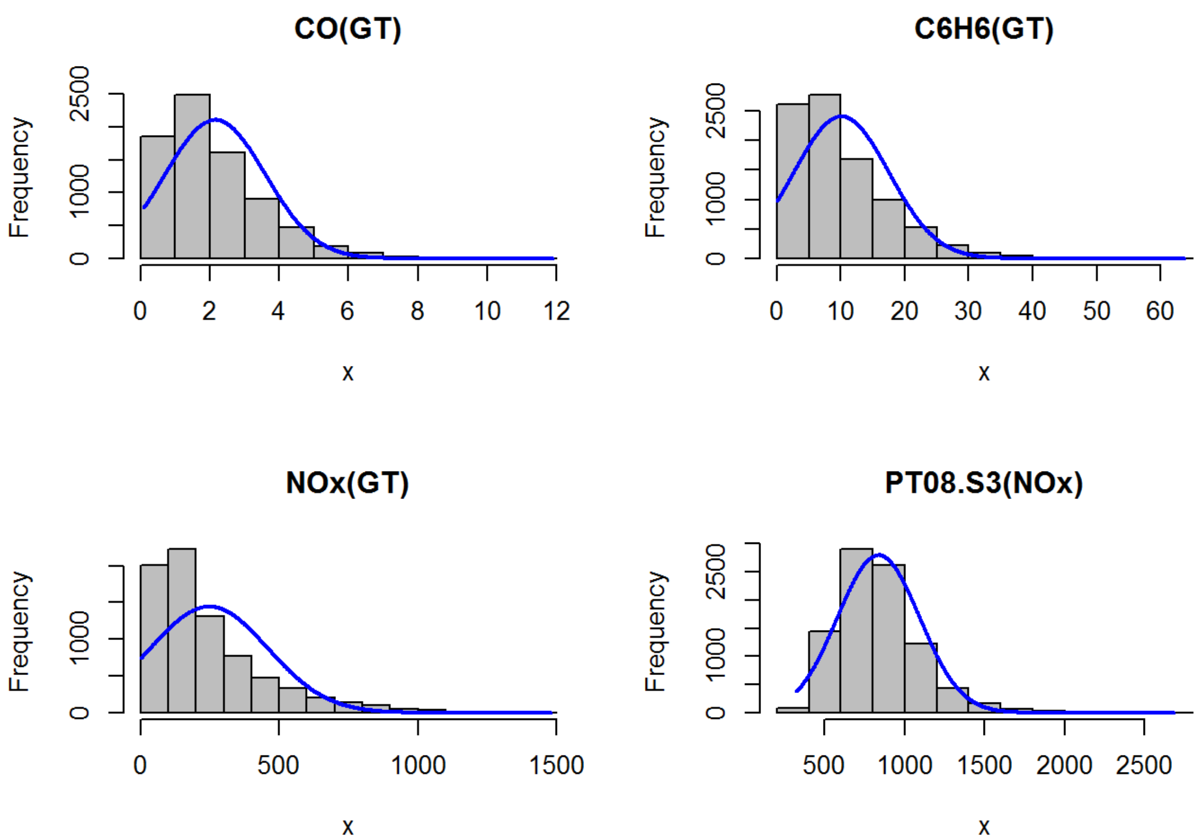
Relevant Summary Statistics

From the results below, a few variables such as `CO(GT)`, `C6H6(GT)`, and `NOx(GT)` are highly skewed.

```
# Checking the normality of the variables of interest
sapply(AirQualityUCI[,c(5:17)], describe)
```

```
##          CO(GT)     PT08.S1(CO) NMHC(GT) C6H6(GT)   PT08.S2(NMHC) NOx(GT)
## vars     1          1           1        1          1             1
## n        7674       8991        914      8991       8991          7718
## mean     2.15275    1099.833    218.8118 10.08311   939.1534      246.8967
## sd       1.453252   217.08      204.4599 7.44982    266.8314      212.9792
## median   1.8        1063        150      8.2        909           180
## trimmed  1.970586   1082.201    183.4631 9.111178   923.2315      211.2536
## mad      1.18608    210.5292    139.3644 6.52344    278.7288      148.26
## min      0.1        647         7        0.1        383           2
## max      11.9       2040        1189     63.7       2214          1479
## range    11.8       1393        1182     63.6       1831          1477
## skew     1.369217   0.7556552   1.55191  1.361078   0.5613786     1.715114
## kurtosis 2.663783   0.3335334   2.239847 2.485434   0.06186024    3.397495
## se       0.01658938 2.289369    6.762933 0.07856729 2.814058      2.424291
##          PT08.S3(NOx) NO2(GT)   PT08.S4(NO2) PT08.S5(O3) T
## vars     1            1         1            1           1
## n        8991         7715      8991         8991        8991
## mean     835.4936     113.0913  1456.265     1022.906    18.31783
## sd       256.8173     48.37011  346.2068     398.4843    8.832116
## median   806          109       1463         963         17.8
## trimmed  814.6854     110.1763  1450.904     996.2361    17.97881
## mad      229.803      47.4432   327.6546     386.9586    9.34038
## min      322          2         551          221         -1.9
## max      2683         340       2775         2523        44.6
## range    2361         338       2224         2302        46.5
## skew     1.101362     0.6214726 0.20532      0.627655    0.3092536
## kurtosis 2.67414      0.4630553 0.07662349   0.07721674  -0.4572531
## se       2.708447     0.5506924 3.651166     4.202495    0.09314526
##          RH          AH
## vars     1           1
## n        8991        8991
## mean     49.2342     1.02553
## sd       17.31689    0.4038126
## median   49.6        0.9954
## trimmed  49.29012    1.014298
## mad      19.71858    0.4241719
## min      9.2         0.1847
## max      88.7        2.231
## range    79.5        2.0463
## skew     -0.03791536 0.2513039
## kurtosis -0.819072   -0.5609963
## se       0.1826274   0.004258688
```

```
par(mfrow = c(2,2))
plotNormalHistogram(AirQualityUCI$`CO(GT)`, main = "CO(GT)")
plotNormalHistogram(AirQualityUCI$`C6H6(GT)`, main = "C6H6(GT)")
plotNormalHistogram(AirQualityUCI$`NOx(GT)`, main = "NOx(GT)")
plotNormalHistogram(AirQualityUCI$`PT08.S3(NOx)`, main = "PT08.S3(NOx)")
```

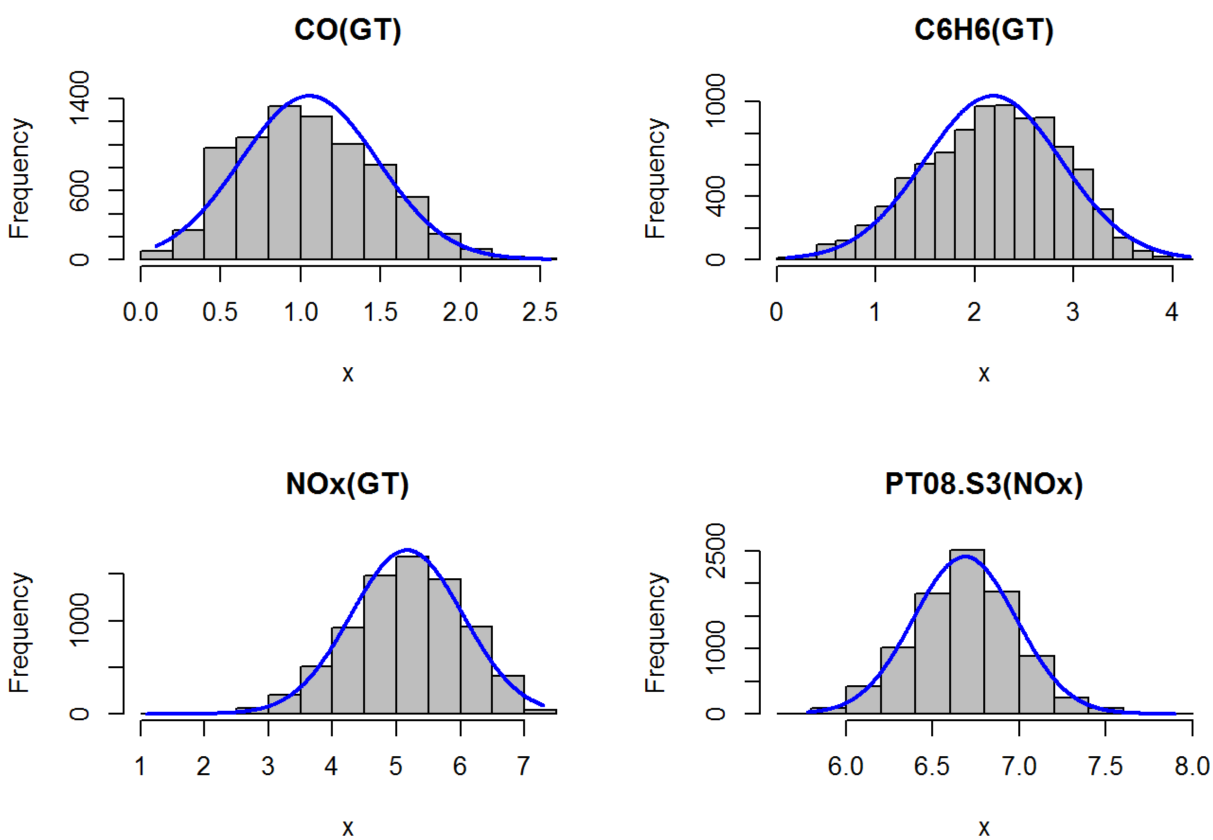


Thus, these variables are normalized before any analysis is conducted. As a result, while the transformed data successfully follow a normal distribution very well, `NOx(GT)` is probably about as close as I can get with this particular data.

```
# Add a value of 1 to each record and log-transform the specific data
AirQualityUCI[,c(5,8,10,11)] <- log(AirQualityUCI[,c(5,8,10,11)]+1)
sapply(AirQualityUCI[,c(5,8,10,11)], describe)
```

```
##          CO(GT)      C6H6(GT)    NOx(GT)     PT08.S3(NOx)
## vars     1           1           1           1
## n        7674        8991        7718        8991
## mean     1.053763    2.182156    5.162929    6.684777
## sd       0.4294719   0.6917609   0.8763291   0.297605
## median   1.029619    2.219203    5.198497    6.693324
## trimmed  1.043119    2.20042     5.186753    6.685411
## mad      0.4527588   0.7444015   0.894567    0.2864501
## min      0.09531018  0.09531018  1.098612    5.777652
## max      2.557227    4.169761    7.299797    7.895063
## range    2.461917    4.074451    6.201185    2.117411
## skew     0.2244873   -0.2330902  -0.2679628  0.02649559
## kurtosis -0.4378552  -0.4342612  -0.1956878  0.1013768
## se       0.004902571 0.007295449 0.009975044 0.003138602
```

```
par(mfrow = c(2,2))
plotNormalHistogram(AirQualityUCI$`CO(GT)`, main = "CO(GT)")
plotNormalHistogram(AirQualityUCI$`C6H6(GT)`, main = "C6H6(GT)")
plotNormalHistogram(AirQualityUCI$`NOx(GT)`, main = "NOx(GT)")
plotNormalHistogram(AirQualityUCI$`PT08.S3(NOx)`, main = "PT08.S3(NOx)")
```

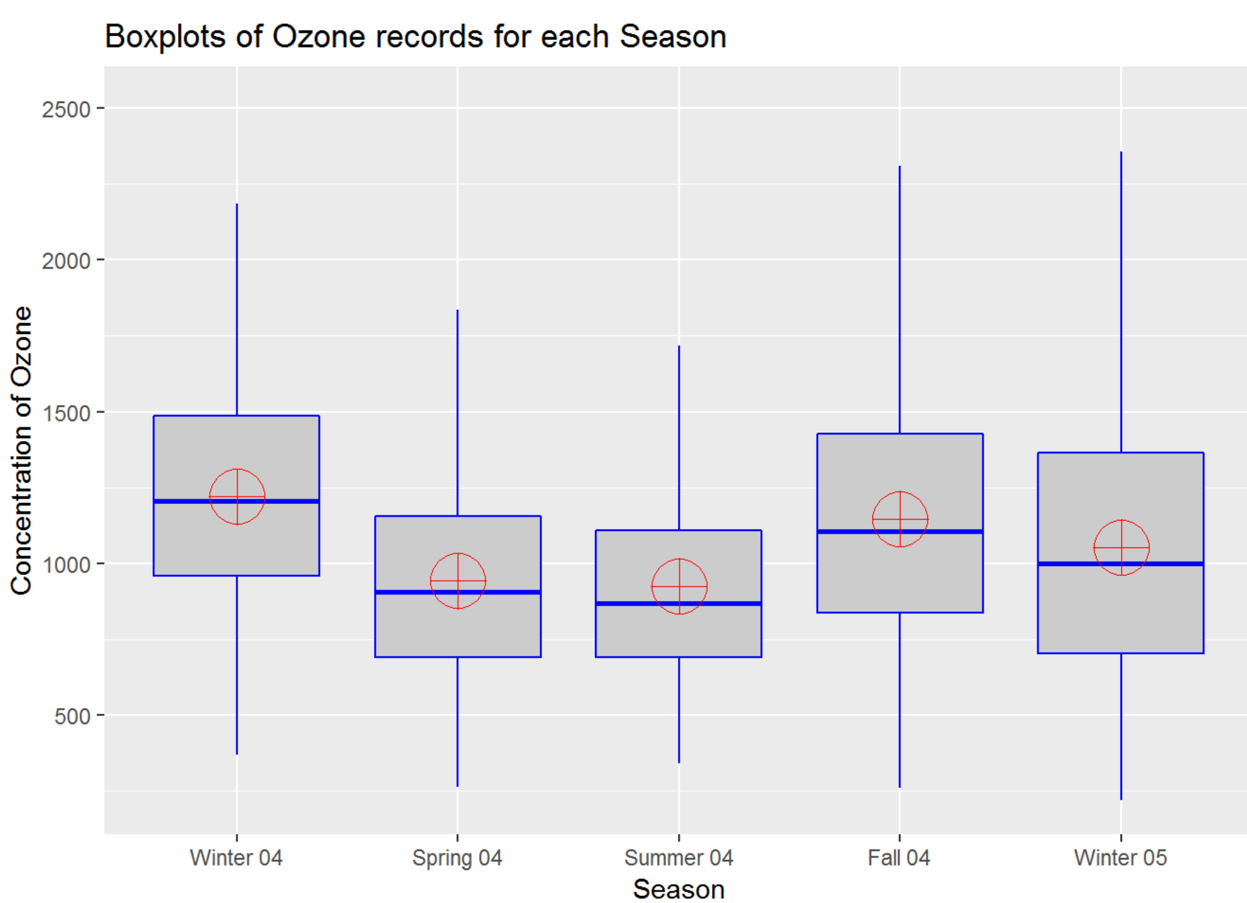


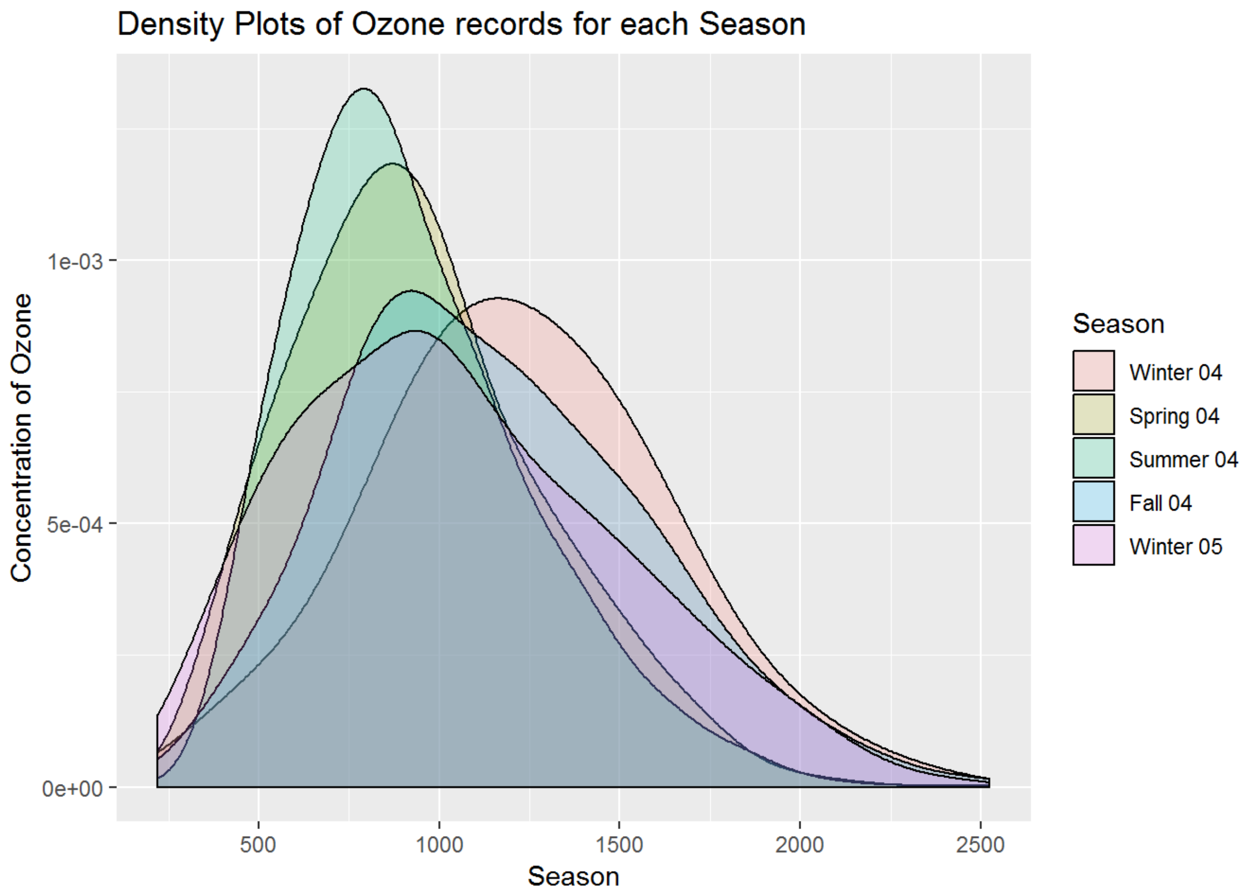

```
par(mfrow = c(1,1))
```

Testing of the assumptions before an ANOVA can be done was conducted for ozone. It is clear that the observations are obtained independently and randomly from the population defined by the factor levels. The data of each factor level are nearly normally distributed and these normal populations have a common variance. In addition, it is apparent that the ozone concentration tends to decrease from Winter to Summer but increases again in Fall.

```
AirQualityUCI$Season <- as.factor(AirQualityUCI$Season)
AirQualityUCI$Season <- factor(AirQualityUCI$Season , levels = levels(AirQualityUCI$Season)[c(
4, 2, 3, 1, 5)])

describeBy(AirQualityUCI$`PT08.S5(O3)`,
           group = AirQualityUCI$Season, mat=TRUE)
```

```
##     item    group1 vars    n      mean       sd median   trimmed      mad
## X11    1 Winter 04    1  222 1220.3378 393.4523   1206 1215.9719 383.9934
## X12    2 Spring 04    1 2157  941.8470 341.8067    906  921.5316 332.1024
## X13    3 Summer 04    1 2112  925.6321 323.5963    868  898.1024 308.3808
## X14    4   Fall 04    1 2108 1147.2694 416.6122   1106 1128.5563 426.9888
## X15    5 Winter 05    1 2392 1053.9678 446.6249   1000 1028.6902 476.6559
##     min  max range      skew    kurtosis        se
## X11 370 2359  1989 0.1685389 -0.08358104 26.406790
## X12 263 2202  1939 0.5482110 -0.01831813  7.359621
## X13 342 2475  2133 0.8317615  0.61151334  7.041362
## X14 261 2523  2262 0.4426613 -0.14316509  9.073957
## X15 221 2494  2273 0.4807518 -0.35944464  9.131925
```



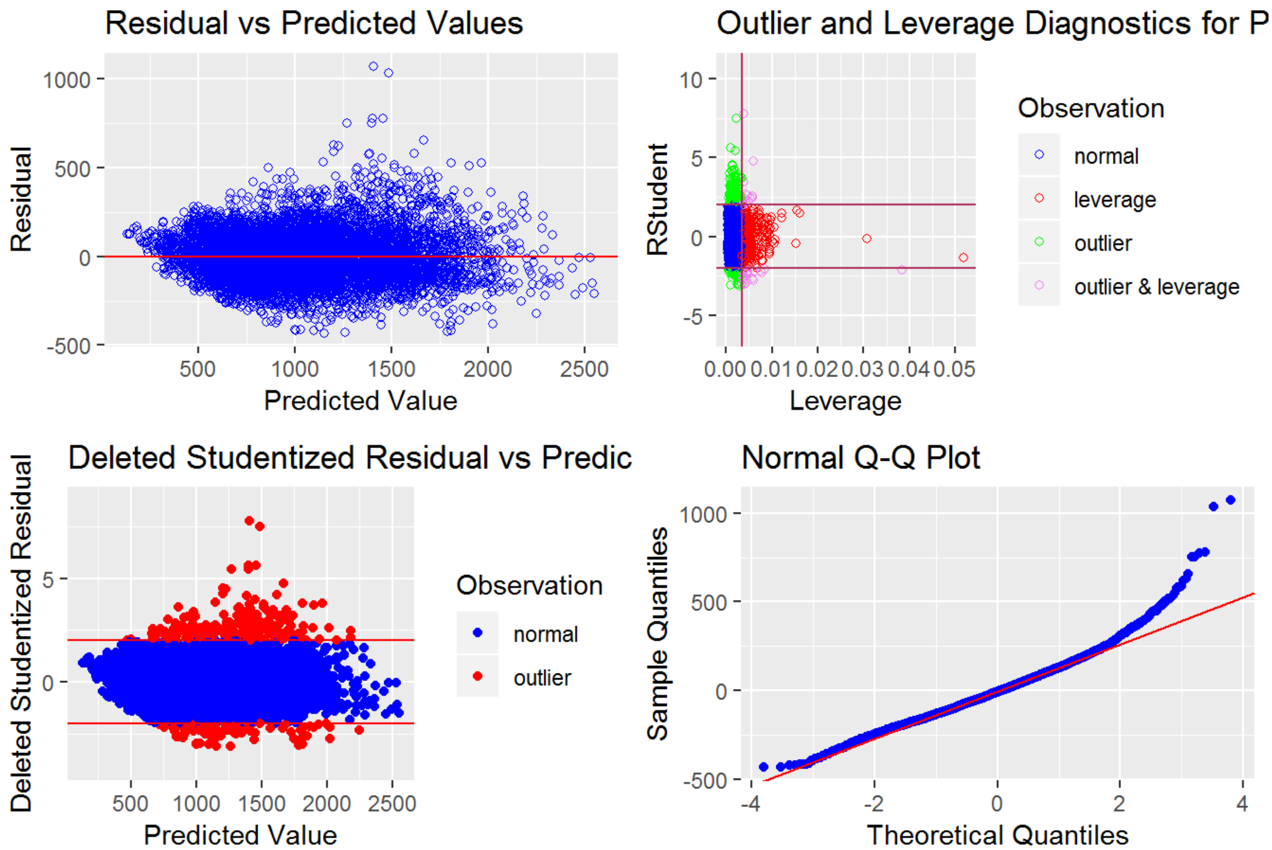


Outliers and Missing Data

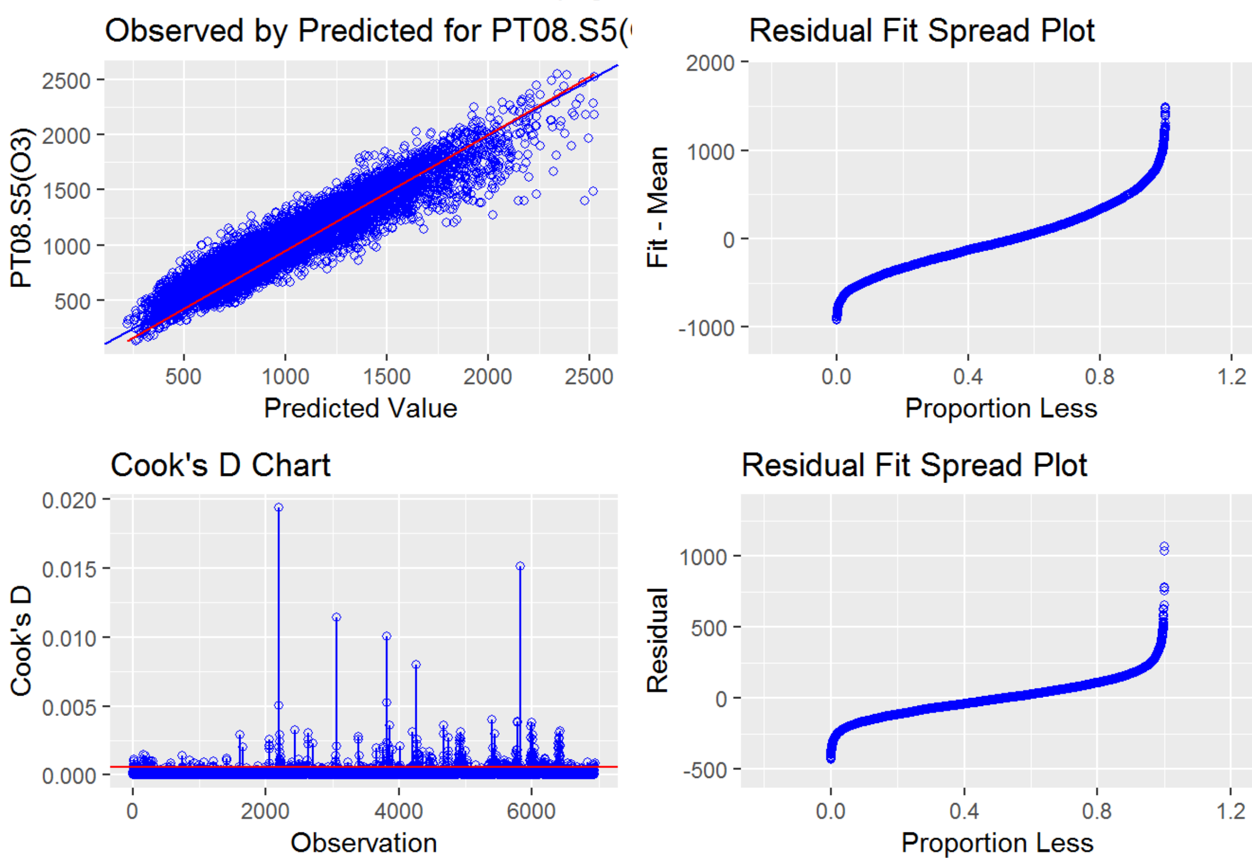
The outlier plots revealed that there are a few extreme values that can influence the analysis. Given that there were initially n = 9358, and the variable `NMHC(GT)` was only reported for 9.7% of the cases, it was therefore excluded from the analyses. Outliers and missing data were corrected by capping it by replacing those observations outside the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of the 95th percentile. It is apparent from the plots of the reduction of the outliers in data from the before and after diagnostic plots.

```
# Outlier Plots
model <- lm(`PT08.S5(O3)` ~ `CO(GT)` + `PT08.S1(CO)` + `C6H6(GT)` + `PT08.S2(NMHC)` + `NOx(GT)
` + `PT08.S3(NOx)` + `NO2(GT)` + `PT08.S4(NO2)` + T + RH + AH, data = AirQualityUCI)
ols_plot_diagnostics(model)
```

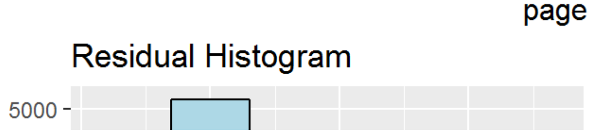
page 1 of 3

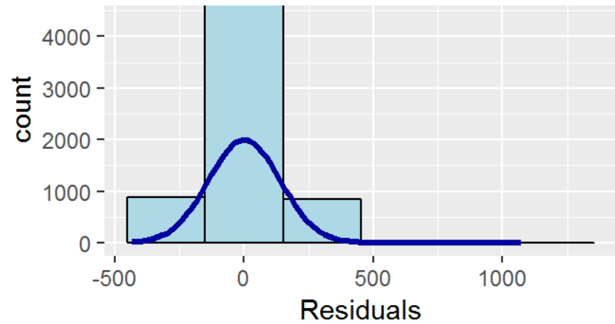


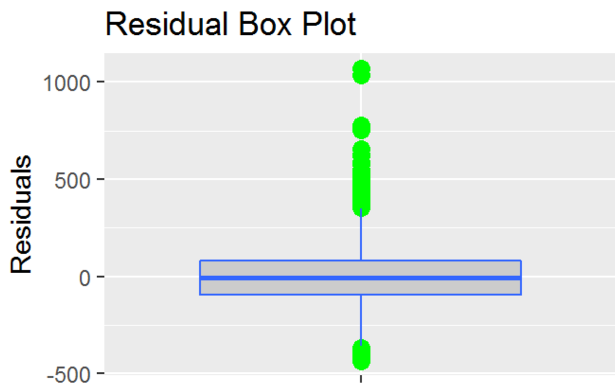
page 2 of 3



page 3 of 3

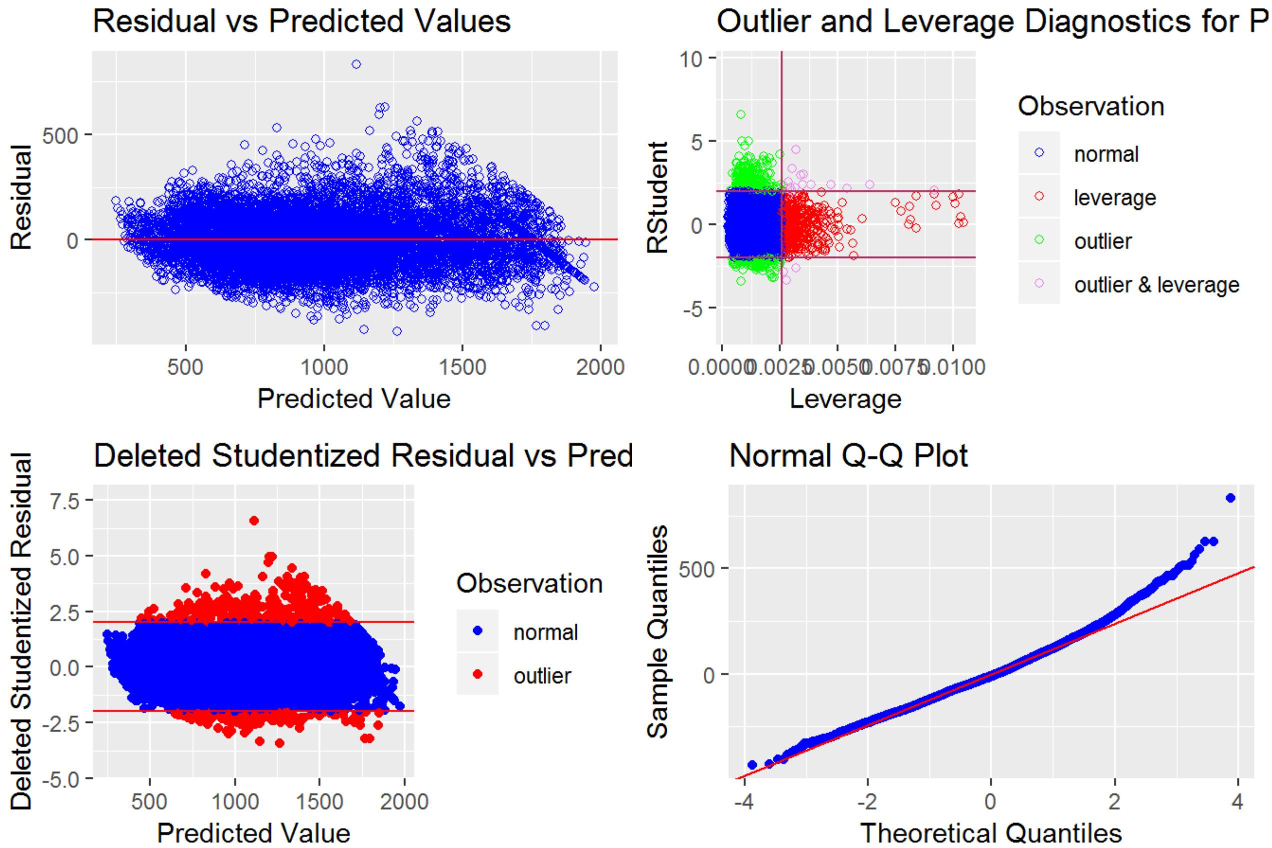




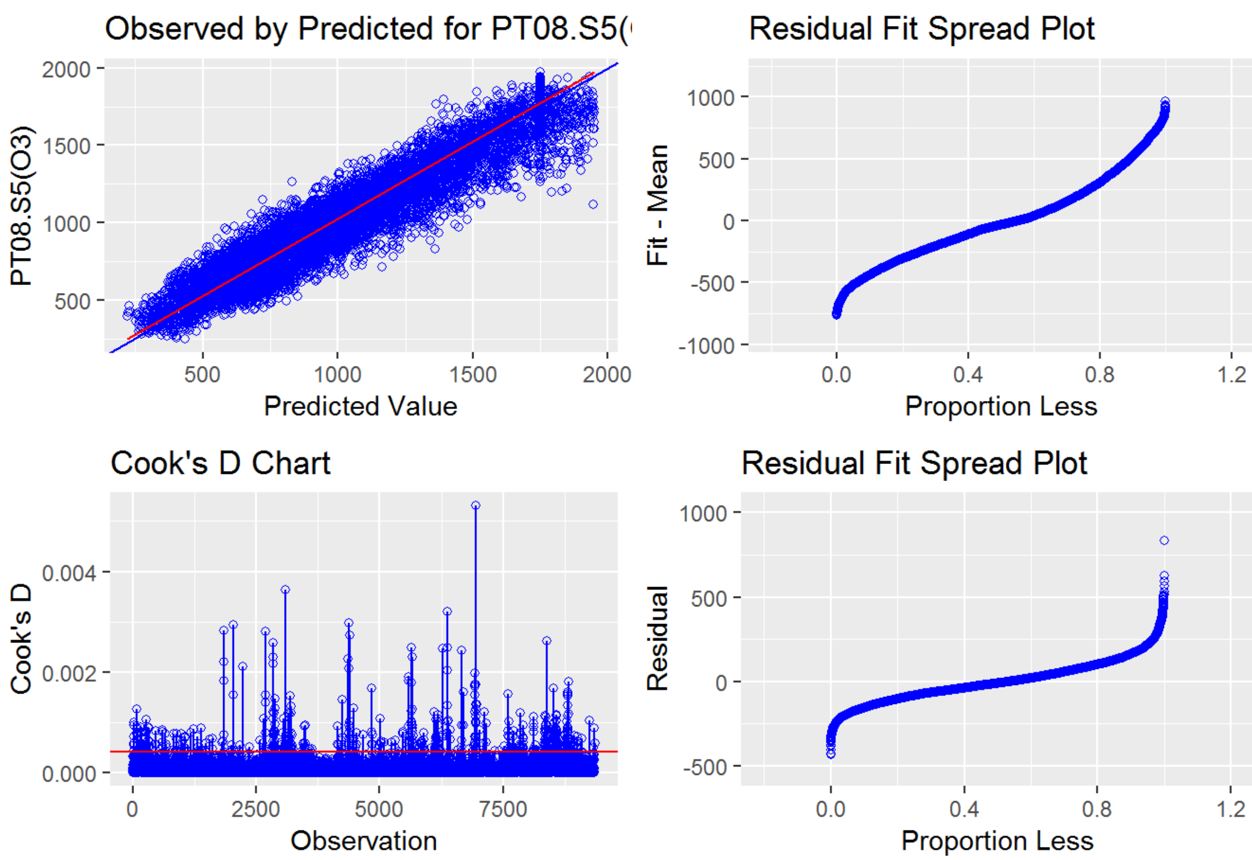



```
AirQualityUCIx <- AirQualityUCI[,-c(7)]

AirQualityUCIx1 <- AirQualityUCIx[,5:16] %>% dplyr::rename_all(paste0, "a")
AirQualityUCIx[,5:16] <- AirQualityUCIx1 %>% mutate_at(vars(ends_with("a")), funs(ifelse(is.na
(.), median(., na.rm = TRUE),.)))

# Remove Outlier
remove_outliers <- function(x, na.rm = TRUE, ...) {
  qnt <- quantile(x, probs = c(.30, .70), na.rm = na.rm, ...)
  caps <- quantile(x, probs=c(.05, .95), na.rm = T)
  H <- 1.5 * IQR(x, na.rm = na.rm)
  y <- x
  y[x < (qnt[1] - H)] <- caps[1]
  y[x > (qnt[2] + H)] <- caps[2]
  y
}
remove_all_outliers <- function(df){
  df[,sapply(df, is.numeric)] <- lapply(df[,sapply(df, is.numeric)], remove_outliers)
  df
}
AirQualityUCIx <- remove_all_outliers(AirQualityUCIx)

# New diagnostic plots
model <- lm(`PT08.S5(O3)` ~ `CO(GT)` + `PT08.S1(CO)` + `C6H6(GT)` + `PT08.S2(NMHC)` + `NOx(GT)
` + `PT08.S3(NOx)` + `NO2(GT)` + `PT08.S4(NO2)` + T + RH + AH, data = AirQualityUCIx)
ols_plot_diagnostics(model)
```

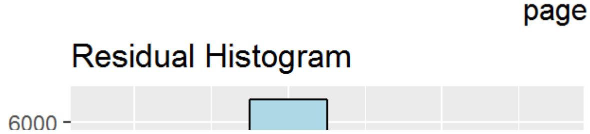
page 1 of 3

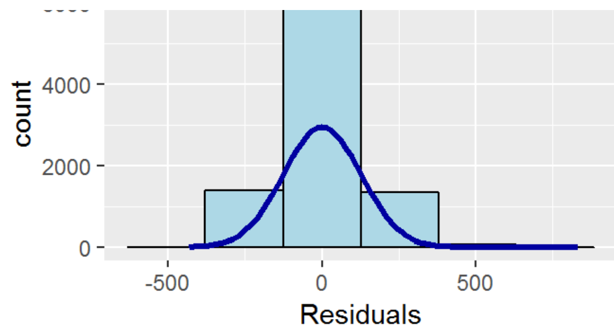


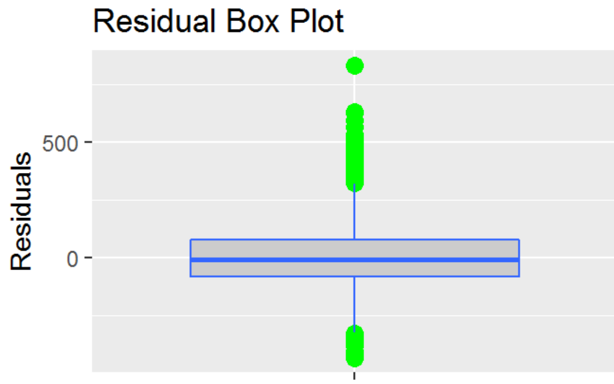
page 2 of 3



page 3 of 3







Initial Tests

Based on the research questions, the statistical test that will be conducted is a regression. Because the data consist of hourly averaged sensor response in addition to the true hourly averaged concentration, there are concerns of possible multicollinearity. Thus, a collinearity diagnostic test is firstly done to examining the diagnostic output for variance inflation factor, tolerance, and Farrar-Glauber F-test. The F-statistic for the variable `PT08.S2(NMHC)` is quite high (44.3239) followed by the variable `C6H6(GT)` (F -value of 31.7851), `T` (F-value of 14.7968), and `PT08.S4(NO2)` (F -value of 14.4250). So the test shows that there are multiple variables that will be the root cause of multicollinearity, specifically the `PT08.S4(NO2)` and `RH` coefficients are non-significant may be due to multicollinearity. Moreover, as expected, there are high partial correlations found to be highly statistically significant. As a solution to deal with multicollinearity, there are several remedial measures such as Stepwise Regression which will be used as a result of this diagnostic test.

```
imcdiag(as.matrix(AirQualityUCIx[,c(5:12,14:16)]), as.matrix(AirQualityUCIx$`PT08.S5(O3)`))
```

```
##
## Call:
## imcdiag(x = as.matrix(AirQualityUCIx[, c(5:12, 14:16)]), y = as.matrix(AirQualityUCIx$`PT0
8.S5(O3)`))
##
##
## All Individual Multicollinearity Diagnostics Result
##
##                   VIF    TOL        Wi        Fi Leamer    CVIF Klein
## CO(GT)         4.2297 0.2364  3018.503  3354.252 0.4862 -0.1253     0
## PT08.S1(CO)    7.4897 0.1335  6065.275  6739.916 0.3654 -0.2219     0
## C6H6(GT)      31.7851 0.0315 28771.776 31972.060 0.1774 -0.9415     1
## PT08.S2(NMHC) 44.3239 0.0226 40490.545 44994.308 0.1502 -1.3129     1
## NOx(GT)        6.7281 0.1486  5353.490  5948.959 0.3855 -0.1993     0
## PT08.S3(NOx)   8.3888 0.1192  6905.570  7673.677 0.3453 -0.2485     0
## NO2(GT)        4.6998 0.2128  3457.870  3842.488 0.4613 -0.1392     0
## PT08.S4(NO2)  14.4250 0.0693 12547.021 13942.626 0.2633 -0.4273     1
## T             14.7968 0.0676 12894.465 14328.716 0.2600 -0.4383     1
## RH             8.5360 0.1172  7043.152  7826.562 0.3423 -0.2528     0
## AH            11.9114 0.0840 10197.750 11332.046 0.2897 -0.3528     1
##
## 1 --> COLLINEARITY is detected by the test
## 0 --> COLLINEARITY is not detected by the test
##
## PT08.S4(NO2) , RH , coefficient(s) are non-significant may be due to multicollinearity
##
## R-square of y on all x: 0.8839
##
## * use method argument to check which regressors may be the reason of collinearity
## ===================================
```

```
corr.test(as.matrix(AirQualityUCIx[,c(5:12,14:16)]), method = "pearson")
```

```
## Call:corr.test(x = as.matrix(AirQualityUCIx[, c(5:12, 14:16)]), method = "pearson")
## Correlation matrix
##                CO(GT) PT08.S1(CO) C6H6(GT) PT08.S2(NMHC) NOx(GT)
## CO(GT)           1.00        0.77     0.80          0.81    0.73
## PT08.S1(CO)      0.77        1.00     0.87          0.89    0.64
## C6H6(GT)         0.80        0.87     1.00          0.98    0.62
## PT08.S2(NMHC)    0.81        0.89     0.98          1.00    0.63
## NOx(GT)          0.73        0.64     0.62          0.63    1.00
## PT08.S3(NOx)    -0.72       -0.84    -0.85         -0.85   -0.70
## NO2(GT)          0.68        0.58     0.58          0.58    0.83
## PT08.S4(NO2)     0.54        0.66     0.75          0.76    0.17
## T                0.05        0.05     0.28          0.25   -0.25
## RH               0.00        0.11    -0.12         -0.10    0.16
## AH               0.04        0.14     0.20          0.19   -0.17
##                PT08.S3(NOx) NO2(GT) PT08.S4(NO2)     T    RH    AH
## CO(GT)                -0.72    0.68         0.54  0.05  0.00  0.04
## PT08.S1(CO)           -0.84    0.58         0.66  0.05  0.11  0.14
## C6H6(GT)              -0.85    0.58         0.75  0.28 -0.12  0.20
## PT08.S2(NMHC)         -0.85    0.58         0.76  0.25 -0.10  0.19
## NOx(GT)               -0.70    0.83         0.17 -0.25  0.16 -0.17
## PT08.S3(NOx)           1.00   -0.61        -0.55 -0.10 -0.09 -0.22
## NO2(GT)               -0.61    1.00         0.14 -0.16 -0.10 -0.30
## PT08.S4(NO2)          -0.55    0.14         1.00  0.58 -0.04  0.65
## T                     -0.10   -0.16         0.58  1.00 -0.58  0.66
## RH                    -0.09   -0.10        -0.04 -0.58  1.00  0.17
## AH                    -0.22   -0.30         0.65  0.66  0.17  1.00
## Sample Size
## [1] 9357
## Probability values (Entries above the diagonal are adjusted for multiple tests.)
##                CO(GT) PT08.S1(CO) C6H6(GT) PT08.S2(NMHC) NOx(GT)
## CO(GT)           0.00           0        0             0       0
## PT08.S1(CO)      0.00           0        0             0       0
## C6H6(GT)         0.00           0        0             0       0
## PT08.S2(NMHC)    0.00           0        0             0       0
## NOx(GT)          0.00           0        0             0       0
## PT08.S3(NOx)     0.00           0        0             0       0
## NO2(GT)          0.00           0        0             0       0
## PT08.S4(NO2)     0.00           0        0             0       0
## T                0.00           0        0             0       0
## RH               0.68           0        0             0       0
## AH               0.00           0        0             0       0
##                PT08.S3(NOx) NO2(GT) PT08.S4(NO2) T   RH AH
## CO(GT)                    0       0            0 0 0.68  0
## PT08.S1(CO)               0       0            0 0 0.00  0
## C6H6(GT)                  0       0            0 0 0.00  0
## PT08.S2(NMHC)             0       0            0 0 0.00  0
## NOx(GT)                   0       0            0 0 0.00  0
## PT08.S3(NOx)              0       0            0 0 0.00  0
## NO2(GT)                   0       0            0 0 0.00  0
## PT08.S4(NO2)              0       0            0 0 0.00  0
## T                         0       0            0 0 0.00  0
```

```
## RH                           0       0            0 0 0.00  0
## AH                           0       0            0 0 0.00  0
##
##  To see confidence intervals of the correlations, print with the short=FALSE option
```

Inference

Part 1

With ground level O_3 being a prominent example of a secondary pollutant with serious consequences to human and Earth, it is a prime indicator of air quality. Since O_3 is formed by the action of sunlight on volatile organic compounds such as Benzene (C_6H_6) in the presence of NO_2, a stepwise variable selection model is conducted to determine what are the predictors that affect the level of ozone in the air. The stepwise variable selection allows variables to be added one at a time to the model, as long as the F-statistic is below the specified α, in this case $\alpha = 0.05$. However, variables already in the model do not necessarily stay in. The steps evaluate all of the variables already included in the model and remove any variable that has an insignificant F-statistic. Only after this test ends, is the best model found, that is when none of the variables can be excluded and every variable included in the model is significant.

Here the dependent variable is the continuous variable, `PT08.S5(O3)`, and the independent variables are the full model to identify the most contributing predictors.

```
# Fit the full model
model <- lm(`PT08.S5(O3)` ~ `CO(GT)` + `PT08.S1(CO)` + `NMHC(GT)` + `C6H6(GT)` + `PT08.S2(NMH
C)` + `NOx(GT)` + `PT08.S3(NOx)` + `NO2(GT)` + `PT08.S4(NO2)` + T + RH + AH + Season, data = A
irQualityUCIx)

# Stepwise Regression model
step <- stepAIC(model, direction = "both")
```

```
# Set up repeated k-fold cross-validation
set.seed(525)

train.control <- trainControl(method = "cv", number = 10)

# Train the model
step.model <- train(`PT08.S5(O3)` ~ `CO(GT)` + `PT08.S1(CO)` + `C6H6(GT)` + `PT08.S2(NMHC)` +
 `NOx(GT)` + `PT08.S3(NOx)` + `NO2(GT)` + `PT08.S4(NO2)` + T + RH + AH + Season, data = AirQua
lityUCIx, method = "lmStepAIC", trControl = train.control, trace = FALSE)

# Model accuracy
step.model$results
```

```
##   parameter     RMSE  Rsquared      MAE   RMSESD  RsquaredSD    MAESD
## 1      none 124.4036 0.8882522 96.91455 3.411205 0.007058904 2.768234
```

```r
# Final model coefficients
step.model$finalModel
```

```
##
## Call:
## lm(formula = .outcome ~ `\`CO(GT)\`` + `\`PT08.S1(CO)\`` + `\`PT08.S2(NMHC)\`` +
##     `\`NOx(GT)\`` + `\`PT08.S3(NOx)\`` + `\`NO2(GT)\`` + T +
##     RH + AH + `SeasonSpring 04` + `SeasonFall 04` + `SeasonWinter 05`,
##     data = dat)
##
## Coefficients:
##          (Intercept)         `\\`CO(GT)\\``    `\\`PT08.S1(CO)\\``
##            2102.6134              -44.3335                 0.6216
## `\\`PT08.S2(NMHC)\\``        `\\`NOx(GT)\\``   `\\`PT08.S3(NOx)\\``
##               0.5776               14.7261              -326.5249
##       `\\`NO2(GT)\\``                      T                     RH
##               0.7402              -11.0144                -0.3885
##                   AH      `SeasonSpring 04`        `SeasonFall 04`
##              38.6434              -57.7141               -54.9683
##     `SeasonWinter 05`
##            -116.1039
```

```r
# Summary of the model
summary(step.model$finalModel)
```

```
##
## Call:
## lm(formula = .outcome ~ `\`CO(GT)\`` + `\`PT08.S1(CO)\`` + `\`PT08.S2(NMHC)\`` +
##     `\`NOx(GT)\`` + `\`PT08.S3(NOx)\`` + `\`NO2(GT)\`` + T +
##     RH + AH + `SeasonSpring 04` + `SeasonFall 04` + `SeasonWinter 05`,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -432.05  -81.90   -7.64   76.32  879.81
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2102.61343  119.95867  17.528  < 2e-16 ***
## `\\`CO(GT)\\``         -44.33351    7.06814  -6.272 3.72e-10 ***
## `\\`PT08.S1(CO)\\``      0.62157    0.01738  35.753  < 2e-16 ***
## `\\`PT08.S2(NMHC)\\``    0.57760    0.01938  29.805  < 2e-16 ***
## `\\`NOx(GT)\\``         14.72608    4.65601   3.163 0.001568 **
## `\\`PT08.S3(NOx)\\``  -326.52487   14.84128 -22.001  < 2e-16 ***
## `\\`NO2(GT)\\``          0.74022    0.07427   9.967  < 2e-16 ***
## T                      -11.01442    0.56953 -19.340  < 2e-16 ***
## RH                      -0.38847    0.21824  -1.780 0.075102 .
## AH                      38.64342   10.62928   3.636 0.000279 ***
## `SeasonSpring 04`      -57.71406    4.31372 -13.379  < 2e-16 ***
## `SeasonFall 04`        -54.96829    4.95452 -11.095  < 2e-16 ***
## `SeasonWinter 05`     -116.10389    6.49671 -17.871  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124.3 on 9344 degrees of freedom
## Multiple R-squared:  0.8884, Adjusted R-squared:  0.8882
## F-statistic:  6197 on 12 and 9344 DF,  p-value: < 2.2e-16
```

After steps, the final model resulted to be:

$$O_3 = 2102.6 - 44.3CO + 0.62CO_{sensor} + 0.58NMHC_{sensor} + 14.73NOx - 326.52NOx_{sensor}$$
$$+0.74NO_2 - 11.01T - 11.01RH + 38.64AH - 38.64Season_{Spring} - 57.71Season_{Fall} - 116.10Season_{Winter}$$

with R^2 = 0.89, suggesting that this model accounts for 89% of the variation in the dependent variable with the independent variables above which is highly impressive. However, when examining the sensor variables and the true concentration variables as two separate models, with the sensors concentrations and full model, the difference in the R^2 was 0.0030852. This suggests that adding these extra variables did not have a very impactful increase on the explained variance. Therefore, it can be safe to select the following model as the best model since it actually only contains these variables, further indicating what the significant predictors of ozone are:

$$O_3 = 2213.9 + 0.63CO_{sensor} + 30.7C_6H_6 + 0.58NMHC_{sensor} - 330.27NOx_{sensor} - 0.049NO_{2sensor}$$
$$-10.03T + 15.94AH - 64.0Season_{Spring} - 56.26Season_{Fall} - 56.26Season_{Winter}$$

```
set.seed(10)

step.model <- train(`PT08.S5(O3)` ~ `CO(GT)` + `C6H6(GT)` + `NOx(GT)` + `NO2(GT)` + T + RH + A
H + Season, data = AirQualityUCIx, method = "lmStepAIC", trControl = train.control, trace = FA
LSE)

# Model accuracy
step.model$results
```

```
##   parameter    RMSE  Rsquared      MAE   RMSESD RsquaredSD    MAESD
## 1      none 145.819 0.8466136 115.8798 4.976303 0.00890242 3.833963
```

```
step.model <- train(`PT08.S5(O3)` ~ `PT08.S1(CO)` + `C6H6(GT)` + `PT08.S2(NMHC)` + `PT08.S3(NO
x)` + `PT08.S4(NO2)` + T + RH + AH + Season, data = AirQualityUCIx, method = "lmStepAIC", trCo
ntrol = train.control, trace = FALSE)

# Model accuracy
step.model$results
```

```
##   parameter     RMSE Rsquared      MAE   RMSESD  RsquaredSD    MAESD
## 1      none 126.0677 0.885167 98.65016 3.788712 0.005580082 2.686454
```

Part 2

During the day, ozone formation occurs. However, during the night, when solar radiation and temperatures are low ozone is destroyed. Similar sequences of reactions occur on an annual basis, with chemical destruction of ozone reaching a peak in winter and a minimum in summer due to variations in sunlight and UV radiation between the seasons. As a result, ozone concentrations tend to be higher in June, July, and August in the northern hemisphere. With these significant predictors that affect the levels of ozone in the air, an analysis of variance (ANOVA) is carried out to understand how do they vary based on the season, and whether any season is more prone to more emissions of one or more of a specific air pollutant than another.

Here the dependent variable is the continuous variable, `PT08.S5(O3)` , and the independent variables is `Season` . In the exploratory data analysis, the assumption for conducting an ANOVA was conducted and passed, thus the testing hypothesis is:

H_0 : the means of the different groups are the same, $\mu_1 = \mu_2 = \ldots = \mu_n$.

H_1 : at least one sample mean is not equal to the others $\mu_j \neq \mu_k$.

As the p-value is less than the significance level 0.05, it can be concluded that there are significant differences between `Season` . The computed Tukey HSD (Tukey Honest Significant Differences) for performing multiple pairwise-comparison between the means of Spring 2004 and Summer 2004 shows that was no significant difference since the adjusted p-value = 0.63. This suggests the concentration of ozone from Spring 2004 to Summer 2004 did not significantly differ, while every other season's ozone concentration did differ.

```
res.aov <- aov(`PT08.S5(O3)`~ Season, data = AirQualityUCIx)

# Summary of the analysis
summary(res.aov)
```

```
##               Df    Sum Sq  Mean Sq F value Pr(>F)
## Season         4 6.585e+07 16461832   125.4 <2e-16 ***
## Residuals   9352 1.228e+09   131323
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
TukeyHSD(res.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = `PT08.S5(O3)` ~ Season, data = AirQualityUCIx)
##
## $Season
##                            diff        lwr         upr     p adj
## Spring 04-Winter 04 -264.32720 -333.94015 -194.714256 0.0000000
## Summer 04-Winter 04 -279.56633 -349.17928 -209.953386 0.0000000
## Fall 04-Winter 04    -79.25429 -148.90217   -9.606414 0.0163803
## Winter 05-Winter 04 -168.31526 -237.51681  -99.113714 0.0000000
## Summer 04-Spring 04  -15.23913  -44.99539   14.517130 0.6294504
## Fall 04-Spring 04    185.07291  155.23501  214.910803 0.0000000
## Winter 05-Spring 04   96.01194   67.23127  124.792611 0.0000000
## Fall 04-Summer 04    200.31204  170.47414  230.149934 0.0000000
## Winter 05-Summer 04  111.25107   82.47040  140.031742 0.0000000
## Winter 05-Fall 04    -89.06097 -117.92604  -60.195900 0.0000000
```

Moreover, from the analysis, it appears that during warmer weather, ozone concentrations were not as low as expected. The table below depicts the temperature per season which can be compared to the ozone, benzene and nitrogen oxides concentration level. Further research revealed that ozone levels do not always increase with increases in temperature, such as when the ratio of VOCs to NOx is low. And, it shows that the NOx and benzene concentrations were lower in the Summer of 2004 even though the temperature was high.

```
temp <- by(AirQualityUCIx$T, AirQualityUCIx$Season, mean)
ozone <- by(AirQualityUCIx$`PT08.S5(O3)`, AirQualityUCIx$Season, mean)
NOx <- by(AirQualityUCIx$`PT08.S3(NOx)`, AirQualityUCIx$Season, mean)
C6H6 <- by(AirQualityUCIx$`C6H6(GT)`, AirQualityUCIx$Season, mean)
cbind(temp, ozone, NOx, C6H6)
```

```
##                temp     ozone      NOx     C6H6
## Winter 04 15.10676 1205.3784 6.852801 2.382551
## Spring 04 19.29900  941.0512 6.825023 2.193859
## Summer 04 28.03098  925.8120 6.689765 2.196847
## Fall 04   16.85362 1126.1241 6.593026 2.360430
## Winter 05 10.37041 1037.0631 6.617649 2.002787
```

Part 3

When non-metallic hydrocarbon are combusted, they produce CO. With the limited data on NMHC concentration (90% missing values), is NMHC still a contributor in predicting the level of CO in the air given this data?

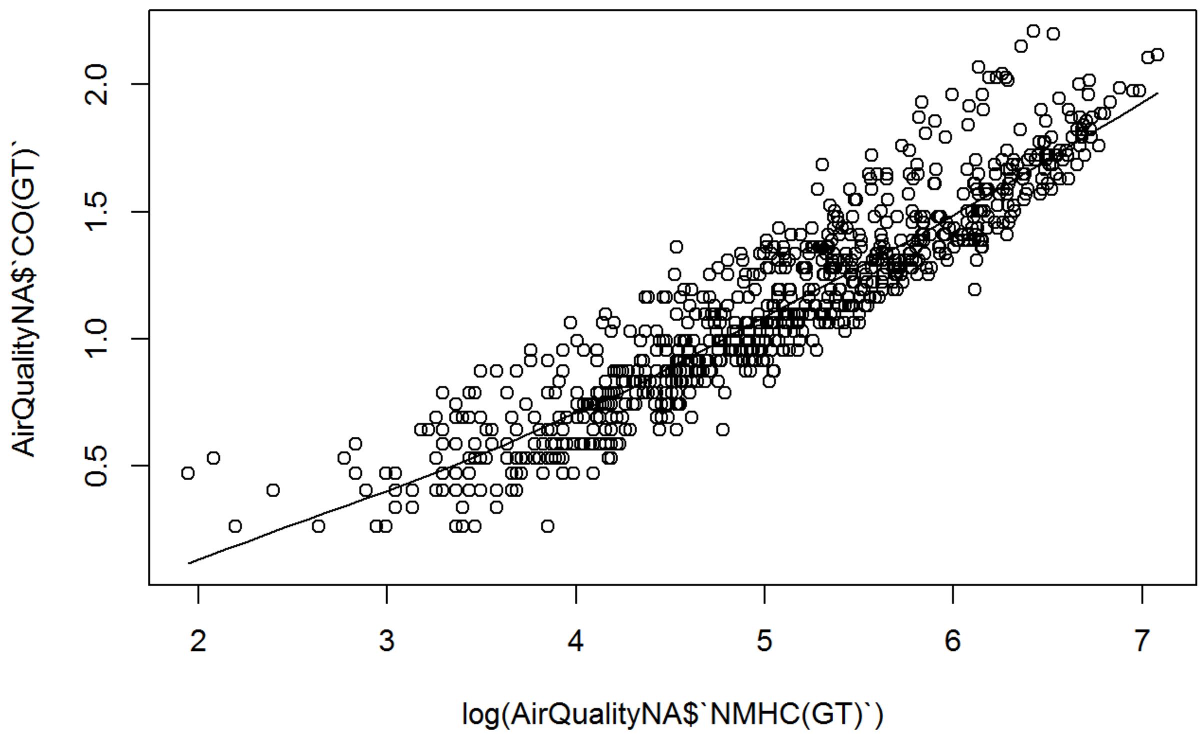
After purging all the incomplete records, the data set has a sample size of n = 827. The assumption for linear regression, as the scatter plot reveals, there is a linear relationship between the dependent variable, `CO(GT)` and independent variable, `NMHC(GT)`. The testing hypothesis in this linear regression becomes:

H_0: the coefficients associated with the variables is equal to zero.

H_A: the coefficients are not equal to zero.

Moreover, the diagnostic plots are used checks for heteroscedasticity, normality, and influential observations.

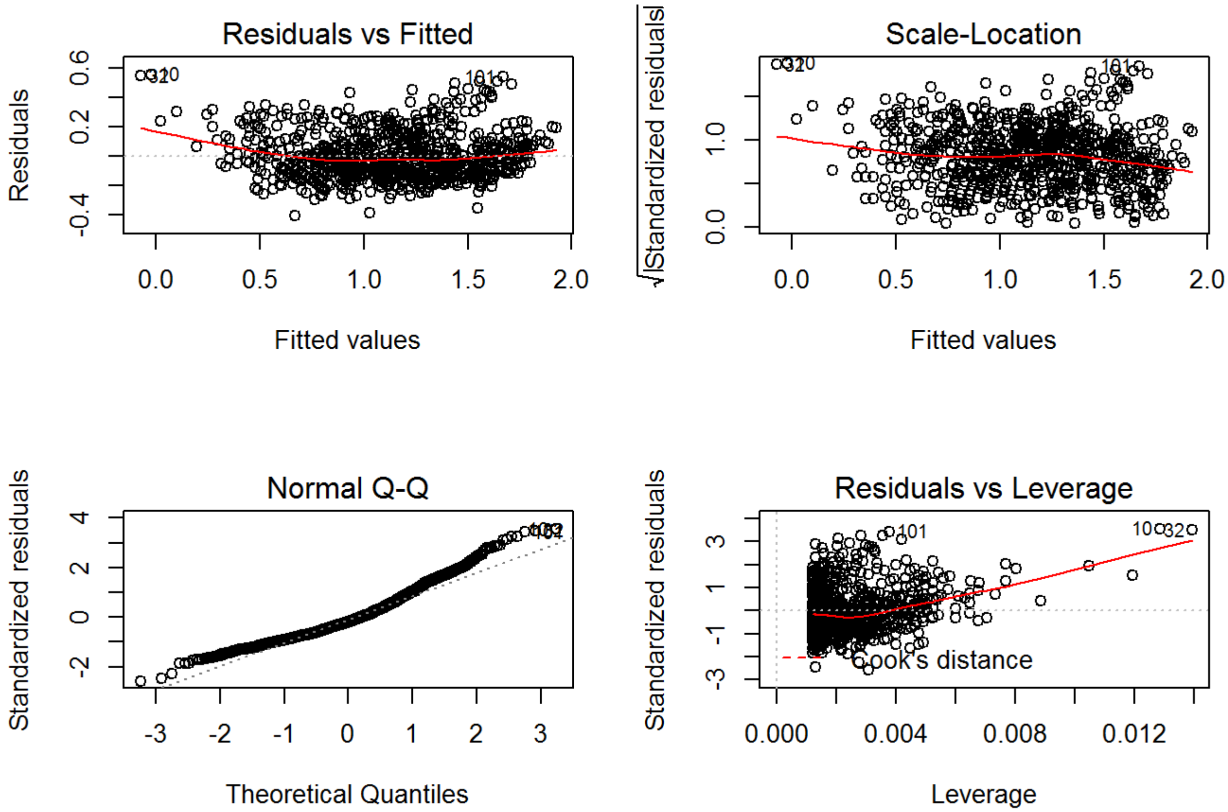
```
# Purge incomplete records
AirQualityNA <- na.omit(AirQualityUCI)

# Test the assumpution for a regression
scatter.smooth(log(AirQualityNA$`NMHC(GT)`) , AirQualityNA$`CO(GT)`)
```




```
fit <- lm(`CO(GT)` ~ log(`NMHC(GT)`), data = AirQualityNA)
summary(fit)
```

```
##
## Call:
## lm(formula = `CO(GT)` ~ log(`NMHC(GT)`), data = AirQualityNA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40471 -0.11215 -0.03053  0.08761  0.55316
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -0.83237    0.02930  -28.41   <2e-16 ***
## log(`NMHC(GT)`)  0.38945    0.00572   68.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1566 on 825 degrees of freedom
## Multiple R-squared:  0.8489, Adjusted R-squared:  0.8488
## F-statistic:  4636 on 1 and 825 DF,  p-value: < 2.2e-16
```

```
layout(matrix(c(1,2,3,4),2,2))
plot(fit)
```



From the model above, it is shows that carbon monoxide constant level in 2004 was -0.832, and that for every 1 mg/m^3 increase in the log(`NMHC(GT)`) concentration in the air, the concentration of carbon monoxide goes up by 0.38945 $microg/m^3$, which can vary by 0.00572 $microg/m^3$.

Moreover, the Residuals vs Fitted and Spread-Location further confirms the linear relationship and homoscedasticity between CO and NMHC since there is an equal spread of residuals around a horizontal line without distinct patterns. The Normal Q-Q follows a straight line well and there aren't many cases outside of the Cook's distance which can be influential to the regression results. Thus, the equation of the line is:

$$CO = -0.83237 + 0.38945 log(NMHC)$$

Conclusion

The prominent example of a secondary pollutant with serious consequences to human and Earth, ozone, was found to differ by season. Ozone's impact on climate consists primarily of changes in temperature. The more ozone in a given parcel of air, the more heat it retains. From this analysis and further research, it is evident that ozone levels do not always increase with temperature because the ratio of VOCs to NOx can sometimes be low. This was the case in Summer 2004, where the NOx and Benzene concentration were different, i.e. lower, resulting in the lower concentration of ozone. Thus, there were significant differences in its level among the seasons.

Moveover, the best model for ozone concentration that accounts for 88% of the variation in the dependent variable is:

$$O_3 = 2213.9 + 0.63 CO_{sensor} + 30.7 C_6H_6 + 0.58 NMHC_{sensor} - 330.27 NOx_{sensor} - 0.049 NO_{2sensor}$$
$$-10.03T + 15.94AH - 64.0 Season_{Spring} - 56.26 Season_{Fall} - 56.26 Season_{Winter}$$

When looking at another air pollutants that are also dependent on other elements in the air, it was found that non-metallic hydrocarbon is a contributor in predicting the level of CO in the air. The linear model shows that for every 1 mg/m^3 increase in the log(`NMHC(GT)`) concentration in the air, the concentration of carbon monoxide goes up by 0.38945 $microg/m^3$.

In this project, exploration of how to deal with multicollinearity variables was insightful. The easiest way to detect multicollinearity is to examine the correlation between each pair of explanatory variables. When it comes to the analysis, there are several remedial measures to deal with this problem such as Principal Component Regression, Ridge Regression, Stepwise Regression, etc.

Lastly, an additional analysis which can provide more insights when it comes to monitoring air quality is by using time series methods. Time series analysis accounts for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. Since this project was interested in the seasonal concentration, trends can be investigated by looking at periodic fluctuations.

References

- Bruce, Peter, and Andrew Bruce. 2017. Practical Statistics for Data Scientists. O'Reilly Media.
- De Vito, S., Massera, E., Piga, M., Martinotto, L., Di Francia, G. "On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario, Sensors and Actuators B: Chemical" 129.2, 22 February 2008, 750-757, ISSN 0925-4005, Web Link (https://www.sciencedirect.com/science/article/pii/S0925400507007691?via%3Dihub).
- Dua, D. and Graff, C. (2019). UCI Machine Learning Repository (http://archive.ics.uci.edu/ml). Irvine, CA: University of California, School of Information and Computer Science.
- EPA. Criteria Air Pollutants (https://www.epa.gov/criteria-air-pollutants). U. S. Environmental Protection Agency. Accessed 28 April 2019.
- EPA. Summary of the Clean Air Act (https://www.epa.gov/laws-regulations/summary-clean-air-act). 42 U.S.C. §7401 et seq. (1970). The official text of the CAA is available in the United States Code on FDSys, from the US Government Printing Office. Accessed 27 April 2019.
- Peng, Huiping. "Air Quality Prediction by Machine Learning Methods." T. University of British Columbia, 2015. Web. 1 Apr. 2019. Theses and Dissertations (ETDs) 2015.
- Sen, Abhishek & Khan, Indrani & Kundu, Debajyoti & Das, Kousik & Datta, Jayanta. (2017). Ecophysiological evaluation of tree species for biomonitoring of air quality and identification of air pollution-tolerant species. Environmental Monitoring and Assessment. 189. 1-15. 10.1007/s10661-017-5955-x.