R Assignment Template

Amelia

2022-05-10

#### Introduction

Here you will answer some fundamental questions like, but are not limited to: 1. What is the purpose of this assignment? 2. What data will you be using? Where is it from, how was it collected, when etc. 3. Are there any importatntn details about the data? 4. Why is this important?

#

#### Methods

1. What type of statistical test are you using?
2. Why does it make sense for this data?
3. What do you hope to gain out from this methodology?

# run tests

#### Results

1. What are the numeric results of the test?
2. Did everything go how you expected?
3. What do the results mean?
4. Is anything significant?
5. This is where you would but graphs

# report results   
# Pick a dataset of your choosing that is appropriate for MLR. (Should be numerical and not have a lot of missing values). Make sure to comment in your report where you downloaded the dataset.  
#   
# I'm choosing an air quality dataset from the following link: https://archive.ics.uci.edu/ml/datasets/Air+Quality#  
#  
# Attributes  
# 0 Date (DD/MM/YYYY)  
# 1 Time (HH.MM.SS)  
# 2 True hourly averaged concentration CO in mg/m^3 (reference analyzer)  
# 3 PT08.S1 (tin oxide) hourly averaged sensor response (nominally CO targeted)  
# 4 True hourly averaged overall Non Metanic HydroCarbons concentration in microg/m^3 (reference analyzer)  
# 5 True hourly averaged Benzene concentration in microg/m^3 (reference analyzer)  
# 6 PT08.S2 (titania) hourly averaged sensor response (nominally NMHC targeted)  
# 7 True hourly averaged NOx concentration in ppb (reference analyzer)  
# 8 PT08.S3 (tungsten oxide) hourly averaged sensor response (nominally NOx targeted)  
# 9 True hourly averaged NO2 concentration in microg/m^3 (reference analyzer)  
# 10 PT08.S4 (tungsten oxide) hourly averaged sensor response (nominally NO2 targeted)  
# 11 PT08.S5 (indium oxide) hourly averaged sensor response (nominally O3 targeted)  
# 12 Temperature in Â°C  
# 13 Relative Humidity (%)  
# 14 AH Absolute Humidity  
#  
#  
#  
  
# Objective: Find the best multilinear regression model of the dataset of your choosing. ?  
#   
# You must include:  
# 1. Box plot and histogram of the dependent variable  
  
# First we load the libraries and data  
library(data.table)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':  
##   
## between, first, last

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library('ggplot2')  
library('corrplot')

## corrplot 0.92 loaded

library('Rmisc')

## Loading required package: lattice

## Loading required package: plyr

## ------------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
## library(plyr); library(dplyr)

## ------------------------------------------------------------------------------

##   
## Attaching package: 'plyr'

## The following objects are masked from 'package:dplyr':  
##   
## arrange, count, desc, failwith, id, mutate, rename, summarise,  
## summarize

library('MASS')

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

library('car')

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

# Load data (pick either white or red wine)  
dt <- read.csv2("C:\\Users\\jerem\\OneDrive\\Documents\\School\\\_REGIS\\2022-05\_Summer\\MSDS660\\Week3\\AirQualityUCI\\AirQualityUCI.csv", sep = ";")  
  
# Change data to datatable or dataframe   
dt <- as.data.frame(dt)  
  
### Exploratory ###   
# Look at summary to see if there are any NAs or characters in the dataset.   
# If there are NAs and/or characters remove them.  
head(dt)

## Date Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.  
## 1 10/03/2004 18.00.00 2.6 1360 150 11.9 1046  
## 2 10/03/2004 19.00.00 2.0 1292 112 9.4 955  
## 3 10/03/2004 20.00.00 2.2 1402 88 9.0 939  
## 4 10/03/2004 21.00.00 2.2 1376 80 9.2 948  
## 5 10/03/2004 22.00.00 1.6 1272 51 6.5 836  
## 6 10/03/2004 23.00.00 1.2 1197 38 4.7 750  
## NOx.GT. PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH AH X X.1  
## 1 166 1056 113 1692 1268 13.6 48.9 0.7578 NA NA  
## 2 103 1174 92 1559 972 13.3 47.7 0.7255 NA NA  
## 3 131 1140 114 1555 1074 11.9 54.0 0.7502 NA NA  
## 4 172 1092 122 1584 1203 11.0 60.0 0.7867 NA NA  
## 5 131 1205 116 1490 1110 11.2 59.6 0.7888 NA NA  
## 6 89 1337 96 1393 949 11.2 59.2 0.7848 NA NA

nrow(dt)

## [1] 9471

ncol(dt)

## [1] 17

str(dt)

## 'data.frame': 9471 obs. of 17 variables:  
## $ Date : chr "10/03/2004" "10/03/2004" "10/03/2004" "10/03/2004" ...  
## $ Time : chr "18.00.00" "19.00.00" "20.00.00" "21.00.00" ...  
## $ CO.GT. : num 2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...  
## $ PT08.S1.CO. : int 1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...  
## $ NMHC.GT. : int 150 112 88 80 51 38 31 31 24 19 ...  
## $ C6H6.GT. : num 11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.7 ...  
## $ PT08.S2.NMHC.: int 1046 955 939 948 836 750 690 672 609 561 ...  
## $ NOx.GT. : int 166 103 131 172 131 89 62 62 45 -200 ...  
## $ PT08.S3.NOx. : int 1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...  
## $ NO2.GT. : int 113 92 114 122 116 96 77 76 60 -200 ...  
## $ PT08.S4.NO2. : int 1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...  
## $ PT08.S5.O3. : int 1268 972 1074 1203 1110 949 733 730 620 501 ...  
## $ T : num 13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 10.3 ...  
## $ RH : num 48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 60.2 ...  
## $ AH : num 0.758 0.726 0.75 0.787 0.789 ...  
## $ X : logi NA NA NA NA NA NA ...  
## $ X.1 : logi NA NA NA NA NA NA ...

summary(dt)

## Date Time CO.GT. PT08.S1.CO.   
## Length:9471 Length:9471 Min. :-200.00 Min. :-200   
## Class :character Class :character 1st Qu.: 0.60 1st Qu.: 921   
## Mode :character Mode :character Median : 1.50 Median :1053   
## Mean : -34.21 Mean :1049   
## 3rd Qu.: 2.60 3rd Qu.:1221   
## Max. : 11.90 Max. :2040   
## NA's :114 NA's :114   
## NMHC.GT. C6H6.GT. PT08.S2.NMHC. NOx.GT.   
## Min. :-200.0 Min. :-200.000 Min. :-200.0 Min. :-200.0   
## 1st Qu.:-200.0 1st Qu.: 4.000 1st Qu.: 711.0 1st Qu.: 50.0   
## Median :-200.0 Median : 7.900 Median : 895.0 Median : 141.0   
## Mean :-159.1 Mean : 1.866 Mean : 894.6 Mean : 168.6   
## 3rd Qu.:-200.0 3rd Qu.: 13.600 3rd Qu.:1105.0 3rd Qu.: 284.0   
## Max. :1189.0 Max. : 63.700 Max. :2214.0 Max. :1479.0   
## NA's :114 NA's :114 NA's :114 NA's :114   
## PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3.   
## Min. :-200 Min. :-200.00 Min. :-200 Min. :-200.0   
## 1st Qu.: 637 1st Qu.: 53.00 1st Qu.:1185 1st Qu.: 700.0   
## Median : 794 Median : 96.00 Median :1446 Median : 942.0   
## Mean : 795 Mean : 58.15 Mean :1391 Mean : 975.1   
## 3rd Qu.: 960 3rd Qu.: 133.00 3rd Qu.:1662 3rd Qu.:1255.0   
## Max. :2683 Max. : 340.00 Max. :2775 Max. :2523.0   
## NA's :114 NA's :114 NA's :114 NA's :114   
## T RH AH X   
## Min. :-200.000 Min. :-200.00 Min. :-200.0000 Mode:logical   
## 1st Qu.: 10.900 1st Qu.: 34.10 1st Qu.: 0.6923 NA's:9471   
## Median : 17.200 Median : 48.60 Median : 0.9768   
## Mean : 9.778 Mean : 39.49 Mean : -6.8376   
## 3rd Qu.: 24.100 3rd Qu.: 61.90 3rd Qu.: 1.2962   
## Max. : 44.600 Max. : 88.70 Max. : 2.2310   
## NA's :114 NA's :114 NA's :114   
## X.1   
## Mode:logical   
## NA's:9471   
##   
##   
##   
##   
##

which(is.na(dt$Date))

## integer(0)

which(is.na(dt$Time))

## integer(0)

which(is.na(dt$CO.GT.))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

which(is.na(dt$PT08.S1.CO.))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

which(is.na(dt$NMHC.GT.))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

which(is.na(dt$C6H6.GT.))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

which(is.na(dt$PT08.S2.NMHC.))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

which(is.na(dt$N0x.GT.))

## integer(0)

which(is.na(dt$PT08.S3.N0x.))

## integer(0)

which(is.na(dt$N02.GT.))

## integer(0)

which(is.na(dt$PT08.S4.N02.))

## integer(0)

which(is.na(dt$PT08.S5.03.))

## integer(0)

which(is.na(dt$T))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

which(is.na(dt$RH))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

which(is.na(dt$AH))

## [1] 9358 9359 9360 9361 9362 9363 9364 9365 9366 9367 9368 9369 9370 9371 9372  
## [16] 9373 9374 9375 9376 9377 9378 9379 9380 9381 9382 9383 9384 9385 9386 9387  
## [31] 9388 9389 9390 9391 9392 9393 9394 9395 9396 9397 9398 9399 9400 9401 9402  
## [46] 9403 9404 9405 9406 9407 9408 9409 9410 9411 9412 9413 9414 9415 9416 9417  
## [61] 9418 9419 9420 9421 9422 9423 9424 9425 9426 9427 9428 9429 9430 9431 9432  
## [76] 9433 9434 9435 9436 9437 9438 9439 9440 9441 9442 9443 9444 9445 9446 9447  
## [91] 9448 9449 9450 9451 9452 9453 9454 9455 9456 9457 9458 9459 9460 9461 9462  
## [106] 9463 9464 9465 9466 9467 9468 9469 9470 9471

#which(is.na(dt$X))  
#which(is.na(dt$X.1))  
  
  
  
  
sum(is.na(dt$PT08.S1.CO.))

## [1] 114

sum(is.na(dt$NMHC.GT.))

## [1] 114

sum(is.na(dt$PT08.S2.NMHC.))

## [1] 114

sum(is.na(dt$X))

## [1] 9471

sum(is.na(dt$X.1))

## [1] 9471

# It looks like we need to remove the X and X.1 columns as they contain null values only.  
  
dtnona <- dt[-c(16, 17)]  
  
names(dt)

## [1] "Date" "Time" "CO.GT." "PT08.S1.CO."   
## [5] "NMHC.GT." "C6H6.GT." "PT08.S2.NMHC." "NOx.GT."   
## [9] "PT08.S3.NOx." "NO2.GT." "PT08.S4.NO2." "PT08.S5.O3."   
## [13] "T" "RH" "AH" "X"   
## [17] "X.1"

names(dtnona)

## [1] "Date" "Time" "CO.GT." "PT08.S1.CO."   
## [5] "NMHC.GT." "C6H6.GT." "PT08.S2.NMHC." "NOx.GT."   
## [9] "PT08.S3.NOx." "NO2.GT." "PT08.S4.NO2." "PT08.S5.O3."   
## [13] "T" "RH" "AH"

nrow(dtnona)

## [1] 9471

dtnona <- dtnona[complete.cases(dtnona), ]  
names(dtnona)

## [1] "Date" "Time" "CO.GT." "PT08.S1.CO."   
## [5] "NMHC.GT." "C6H6.GT." "PT08.S2.NMHC." "NOx.GT."   
## [9] "PT08.S3.NOx." "NO2.GT." "PT08.S4.NO2." "PT08.S5.O3."   
## [13] "T" "RH" "AH"

nrow(dtnona)

## [1] 9357

which(is.na(dtnona$Date))

## integer(0)

which(is.na(dtnona$Time))

## integer(0)

which(is.na(dtnona$CO.GT.))

## integer(0)

which(is.na(dtnona$PT08.S1.CO.))

## integer(0)

which(is.na(dtnona$NMHC.GT.))

## integer(0)

which(is.na(dtnona$C6H6.GT.))

## integer(0)

which(is.na(dtnona$PT08.S2.NMHC.))

## integer(0)

which(is.na(dtnona$N0x.GT.))

## integer(0)

which(is.na(dtnona$PT08.S3.N0x.))

## integer(0)

which(is.na(dtnona$N02.GT.))

## integer(0)

which(is.na(dtnona$PT08.S4.N02.))

## integer(0)

which(is.na(dtnona$PT08.S5.03.))

## integer(0)

which(is.na(dtnona$T))

## integer(0)

which(is.na(dtnona$RH))

## integer(0)

which(is.na(dtnona$AH))

## integer(0)

# Okay! so we've cleaned all null values. Now I'm going to remove the date column and convert time to a numerical hour  
  
dtfinal <- dtnona[-c(1)]  
names(dtfinal)

## [1] "Time" "CO.GT." "PT08.S1.CO." "NMHC.GT."   
## [5] "C6H6.GT." "PT08.S2.NMHC." "NOx.GT." "PT08.S3.NOx."   
## [9] "NO2.GT." "PT08.S4.NO2." "PT08.S5.O3." "T"   
## [13] "RH" "AH"

head(dtfinal)

## Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC. NOx.GT.  
## 1 18.00.00 2.6 1360 150 11.9 1046 166  
## 2 19.00.00 2.0 1292 112 9.4 955 103  
## 3 20.00.00 2.2 1402 88 9.0 939 131  
## 4 21.00.00 2.2 1376 80 9.2 948 172  
## 5 22.00.00 1.6 1272 51 6.5 836 131  
## 6 23.00.00 1.2 1197 38 4.7 750 89  
## PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH AH  
## 1 1056 113 1692 1268 13.6 48.9 0.7578  
## 2 1174 92 1559 972 13.3 47.7 0.7255  
## 3 1140 114 1555 1074 11.9 54.0 0.7502  
## 4 1092 122 1584 1203 11.0 60.0 0.7867  
## 5 1205 116 1490 1110 11.2 59.6 0.7888  
## 6 1337 96 1393 949 11.2 59.2 0.7848

dtfinal$Time = as.numeric(substr(dtfinal$Time, start = 1, stop = 2))  
head(dtfinal)

## Time CO.GT. PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx.  
## 1 18 2.6 1360 150 11.9 1046 166 1056  
## 2 19 2.0 1292 112 9.4 955 103 1174  
## 3 20 2.2 1402 88 9.0 939 131 1140  
## 4 21 2.2 1376 80 9.2 948 172 1092  
## 5 22 1.6 1272 51 6.5 836 131 1205  
## 6 23 1.2 1197 38 4.7 750 89 1337  
## NO2.GT. PT08.S4.NO2. PT08.S5.O3. T RH AH  
## 1 113 1692 1268 13.6 48.9 0.7578  
## 2 92 1559 972 13.3 47.7 0.7255  
## 3 114 1555 1074 11.9 54.0 0.7502  
## 4 122 1584 1203 11.0 60.0 0.7867  
## 5 116 1490 1110 11.2 59.6 0.7888  
## 6 96 1393 949 11.2 59.2 0.7848

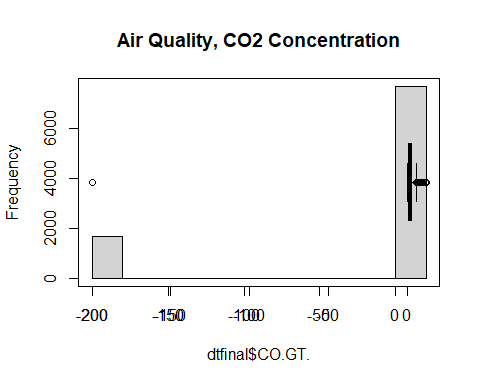
summary(dtfinal)

## Time CO.GT. PT08.S1.CO. NMHC.GT.   
## Min. : 0.0 Min. :-200.00 Min. :-200 Min. :-200.0   
## 1st Qu.: 5.0 1st Qu.: 0.60 1st Qu.: 921 1st Qu.:-200.0   
## Median :11.0 Median : 1.50 Median :1053 Median :-200.0   
## Mean :11.5 Mean : -34.21 Mean :1049 Mean :-159.1   
## 3rd Qu.:18.0 3rd Qu.: 2.60 3rd Qu.:1221 3rd Qu.:-200.0   
## Max. :23.0 Max. : 11.90 Max. :2040 Max. :1189.0   
## C6H6.GT. PT08.S2.NMHC. NOx.GT. PT08.S3.NOx.   
## Min. :-200.000 Min. :-200.0 Min. :-200.0 Min. :-200   
## 1st Qu.: 4.000 1st Qu.: 711.0 1st Qu.: 50.0 1st Qu.: 637   
## Median : 7.900 Median : 895.0 Median : 141.0 Median : 794   
## Mean : 1.866 Mean : 894.6 Mean : 168.6 Mean : 795   
## 3rd Qu.: 13.600 3rd Qu.:1105.0 3rd Qu.: 284.0 3rd Qu.: 960   
## Max. : 63.700 Max. :2214.0 Max. :1479.0 Max. :2683   
## NO2.GT. PT08.S4.NO2. PT08.S5.O3. T   
## Min. :-200.00 Min. :-200 Min. :-200.0 Min. :-200.000   
## 1st Qu.: 53.00 1st Qu.:1185 1st Qu.: 700.0 1st Qu.: 10.900   
## Median : 96.00 Median :1446 Median : 942.0 Median : 17.200   
## Mean : 58.15 Mean :1391 Mean : 975.1 Mean : 9.778   
## 3rd Qu.: 133.00 3rd Qu.:1662 3rd Qu.:1255.0 3rd Qu.: 24.100   
## Max. : 340.00 Max. :2775 Max. :2523.0 Max. : 44.600   
## RH AH   
## Min. :-200.00 Min. :-200.0000   
## 1st Qu.: 34.10 1st Qu.: 0.6923   
## Median : 48.60 Median : 0.9768   
## Mean : 39.49 Mean : -6.8376   
## 3rd Qu.: 61.90 3rd Qu.: 1.2962   
## Max. : 88.70 Max. : 2.2310

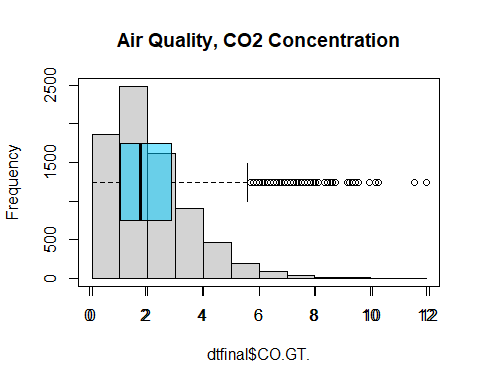
str(dtfinal)

## 'data.frame': 9357 obs. of 14 variables:  
## $ Time : num 18 19 20 21 22 23 0 1 2 3 ...  
## $ CO.GT. : num 2.6 2 2.2 2.2 1.6 1.2 1.2 1 0.9 0.6 ...  
## $ PT08.S1.CO. : int 1360 1292 1402 1376 1272 1197 1185 1136 1094 1010 ...  
## $ NMHC.GT. : int 150 112 88 80 51 38 31 31 24 19 ...  
## $ C6H6.GT. : num 11.9 9.4 9 9.2 6.5 4.7 3.6 3.3 2.3 1.7 ...  
## $ PT08.S2.NMHC.: int 1046 955 939 948 836 750 690 672 609 561 ...  
## $ NOx.GT. : int 166 103 131 172 131 89 62 62 45 -200 ...  
## $ PT08.S3.NOx. : int 1056 1174 1140 1092 1205 1337 1462 1453 1579 1705 ...  
## $ NO2.GT. : int 113 92 114 122 116 96 77 76 60 -200 ...  
## $ PT08.S4.NO2. : int 1692 1559 1555 1584 1490 1393 1333 1333 1276 1235 ...  
## $ PT08.S5.O3. : int 1268 972 1074 1203 1110 949 733 730 620 501 ...  
## $ T : num 13.6 13.3 11.9 11 11.2 11.2 11.3 10.7 10.7 10.3 ...  
## $ RH : num 48.9 47.7 54 60 59.6 59.2 56.8 60 59.7 60.2 ...  
## $ AH : num 0.758 0.726 0.75 0.787 0.789 ...

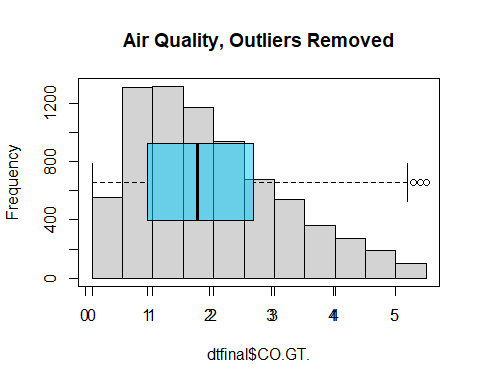
# Awesome, all the data is clean and numeric now  
  
hist(dtfinal$CO.GT., main="Air Quality, CO2 Concentration")  
par(new = TRUE)  
boxplot(dtfinal$CO.GT., horizontal = TRUE, col = rgb(0, 0.8, 1, alpha = 0.5))  
box()



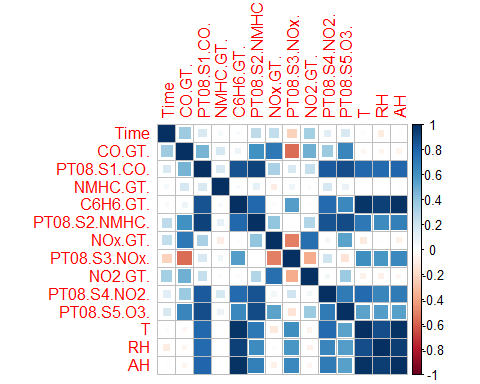
# it looks like we want to remove all the negative concentrations, these are erroneous  
  
dtfinal <- subset(dtfinal, dtfinal$CO.GT. >= 0)  
  
hist(dtfinal$CO.GT., main="Air Quality, CO2 Concentration")  
par(new = TRUE)  
boxplot(dtfinal$CO.GT., horizontal = TRUE, col = rgb(0, 0.8, 1, alpha = 0.5))  
box()



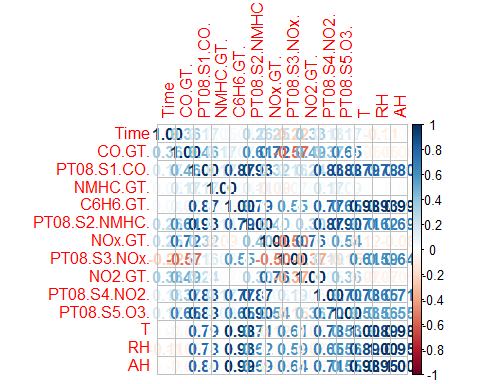
# now let's remove outliers  
  
Q1 <- quantile(dtfinal$CO.GT., .25)  
Q3 <- quantile(dtfinal$CO.GT., .75)  
IQR <- IQR(dtfinal$CO.GT.)  
#only keep rows in dataframe that have values within 1.5\*IQR of Q1 and Q3  
dtfinal <- subset(dtfinal, dtfinal$CO.GT.> (Q1 - 1.5\*IQR) & dtfinal$CO.GT.< (Q3 + 1.5\*IQR))  
  
hist(dtfinal$CO.GT., main="Air Quality, Outliers Removed")  
par(new = TRUE)  
boxplot(dtfinal$CO.GT., horizontal = TRUE, col = rgb(0, 0.8, 1, alpha = 0.5))  
box()



# 2. A correlation plot of all the numerical variables  
  
corrplot(cor(dtfinal), method = 'square')



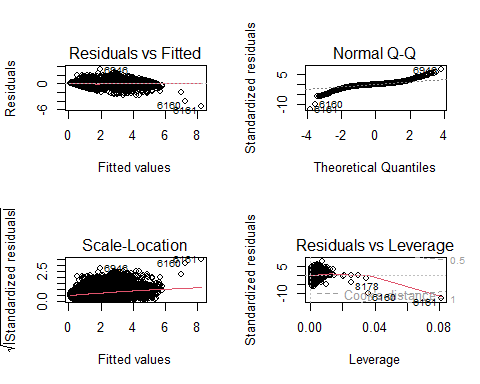
corrplot(cor(dtfinal), method = 'number')



# it looks like "True hourly averaged concentration CO in mg/m^3 (reference analyzer)" is our dependent variable, CO.GT  
  
#Correlations  
#most positive  
# N0x.GT. (0.72), PT08.S5.03. (0.65), PT08.S2.NMHC. (0.61), N02.GT. (0.49), PT08.S1.CO. (0.46), PT08.S4.N02. (0.37), Time (0.36)  
#most negative  
# PT08.S3.N0x. (-0.57)  
  
# 3. A MLR model that has a summary, residual plots, and a VIF analysis  
  
# First, I'm going to create a model that takes every variable  
  
m1 <- lm(CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. + NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH + AH, data = dtfinal)  
summary(m1)

##   
## Call:  
## lm(formula = CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. +   
## PT08.S2.NMHC. + NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. +   
## PT08.S5.O3. + T + RH + AH, data = dtfinal)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1390 -0.1792 -0.0043 0.1796 3.2225   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.159e-01 1.419e-01 -5.045 4.65e-07 \*\*\*  
## Time 1.310e-02 8.707e-04 15.046 < 2e-16 \*\*\*  
## PT08.S1.CO. 9.188e-04 7.303e-05 12.580 < 2e-16 \*\*\*  
## NMHC.GT. 4.496e-04 4.487e-05 10.019 < 2e-16 \*\*\*  
## C6H6.GT. 6.789e-02 5.551e-03 12.231 < 2e-16 \*\*\*  
## PT08.S2.NMHC. 9.183e-04 1.779e-04 5.163 2.49e-07 \*\*\*  
## NOx.GT. 2.085e-03 5.985e-05 34.837 < 2e-16 \*\*\*  
## PT08.S3.NOx. -1.481e-05 4.401e-05 -0.337 0.736483   
## NO2.GT. -2.024e-03 1.058e-04 -19.128 < 2e-16 \*\*\*  
## PT08.S4.NO2. 4.957e-04 5.258e-05 9.427 < 2e-16 \*\*\*  
## PT08.S5.O3. -1.333e-04 3.805e-05 -3.502 0.000465 \*\*\*  
## T -2.029e-02 1.656e-03 -12.252 < 2e-16 \*\*\*  
## RH -5.657e-03 6.549e-04 -8.638 < 2e-16 \*\*\*  
## AH -5.719e-02 5.935e-03 -9.636 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4343 on 7418 degrees of freedom  
## Multiple R-squared: 0.8698, Adjusted R-squared: 0.8696   
## F-statistic: 3812 on 13 and 7418 DF, p-value: < 2.2e-16

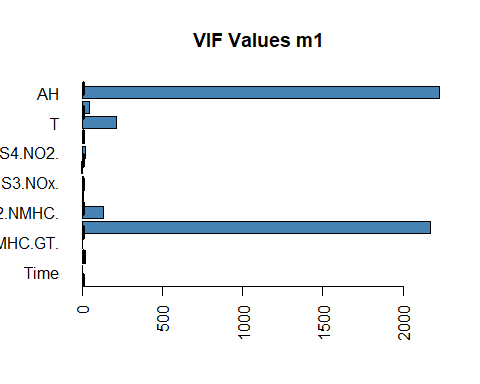
par(mfrow = c(2, 2))  
plot(m1)



vif\_values1 <- vif(m1)  
vif\_values1

## Time PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.   
## 1.434788 22.037758 1.665556 2168.050549 133.897901   
## NOx.GT. PT08.S3.NOx. NO2.GT. PT08.S4.NO2. PT08.S5.O3.   
## 6.113763 7.896918 3.147522 23.068118 11.142512   
## T RH AH   
## 211.946324 46.545034 2228.230225

par(mfrow = c(1,1))  
barplot(vif\_values1, main= "VIF Values m1", horiz = TRUE, col = "steelblue", las=2)  
abline(v = 10, lwd = 3, lty = 2)



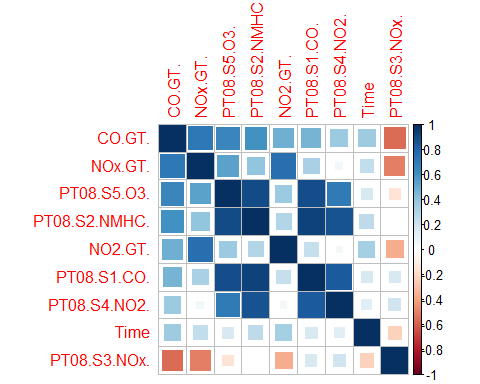
# Next, I want to create a model that just takes the most highly correlated variables with CO.GT, mentioned above  
names(dtfinal)

## [1] "Time" "CO.GT." "PT08.S1.CO." "NMHC.GT."   
## [5] "C6H6.GT." "PT08.S2.NMHC." "NOx.GT." "PT08.S3.NOx."   
## [9] "NO2.GT." "PT08.S4.NO2." "PT08.S5.O3." "T"   
## [13] "RH" "AH"

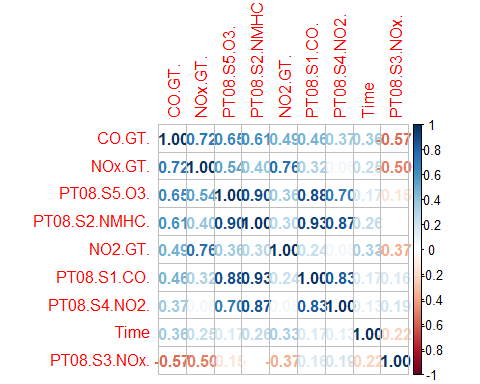
dtm2 <- dtfinal[ , c('CO.GT.','NOx.GT.', 'PT08.S5.O3.', 'PT08.S2.NMHC.', 'NO2.GT.', 'PT08.S1.CO.', 'PT08.S4.NO2.', 'Time', 'PT08.S3.NOx.')]  
names(dtm2)

## [1] "CO.GT." "NOx.GT." "PT08.S5.O3." "PT08.S2.NMHC."  
## [5] "NO2.GT." "PT08.S1.CO." "PT08.S4.NO2." "Time"   
## [9] "PT08.S3.NOx."

corrplot(cor(dtm2), method = 'square')



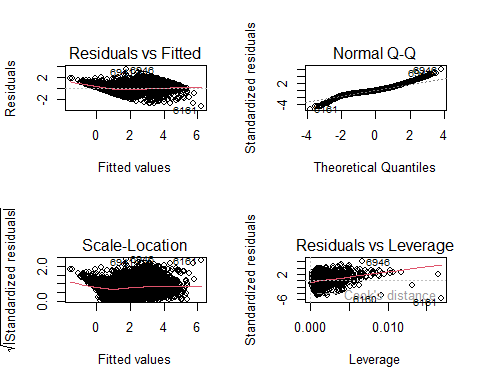
corrplot(cor(dtm2), method = 'number')



m2 <- lm(CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S5.O3. + PT08.S2.NMHC. + PT08.S1.CO. + PT08.S4.NO2. + Time + PT08.S3.NOx., data = dtm2)  
summary(m2)

##   
## Call:  
## lm(formula = CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S5.O3. + PT08.S2.NMHC. +   
## PT08.S1.CO. + PT08.S4.NO2. + Time + PT08.S3.NOx., data = dtm2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1073 -0.3539 -0.0863 0.3022 3.4165   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.157e+00 3.209e-02 36.068 < 2e-16 \*\*\*  
## NO2.GT. -2.047e-03 1.262e-04 -16.217 < 2e-16 \*\*\*  
## NOx.GT. 2.633e-03 6.671e-05 39.474 < 2e-16 \*\*\*  
## PT08.S5.O3. 1.139e-04 4.898e-05 2.325 0.0201 \*   
## PT08.S2.NMHC. 1.847e-03 8.693e-05 21.246 < 2e-16 \*\*\*  
## PT08.S1.CO. -7.288e-04 7.343e-05 -9.925 < 2e-16 \*\*\*  
## PT08.S4.NO2. 2.578e-04 3.930e-05 6.559 5.77e-11 \*\*\*  
## Time 1.724e-02 1.115e-03 15.461 < 2e-16 \*\*\*  
## PT08.S3.NOx. -1.318e-03 3.046e-05 -43.256 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.578 on 7423 degrees of freedom  
## Multiple R-squared: 0.7692, Adjusted R-squared: 0.7689   
## F-statistic: 3092 on 8 and 7423 DF, p-value: < 2.2e-16

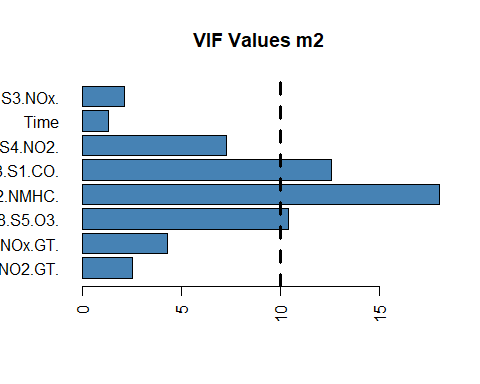
par(mfrow = c(2, 2))  
plot(m2)



vif\_values2 <- vif(m2)  
vif\_values2

## NO2.GT. NOx.GT. PT08.S5.O3. PT08.S2.NMHC. PT08.S1.CO.   
## 2.528600 4.287766 10.419446 18.052733 12.575618   
## PT08.S4.NO2. Time PT08.S3.NOx.   
## 7.271724 1.328498 2.134814

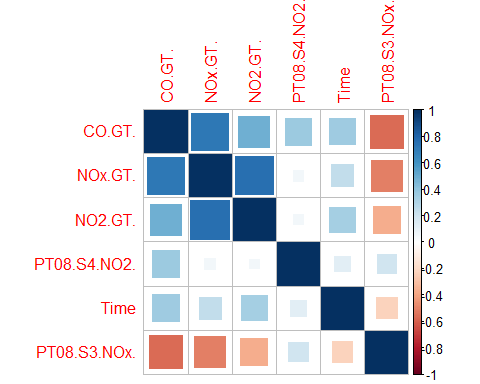
par(mfrow = c(1,1))  
barplot(vif\_values2, main= "VIF Values m2", horiz = TRUE, col = "steelblue", las=2)  
abline(v = 10, lwd = 3, lty = 2)



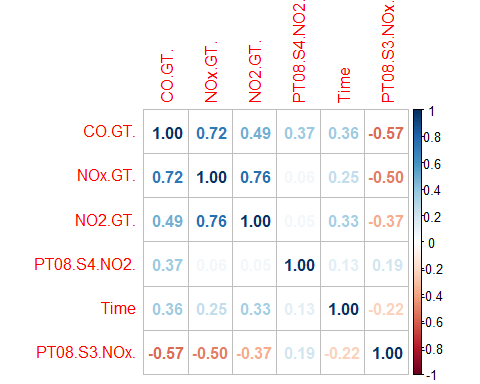
# Now based on the VIF value I will remove PT08.S1.CO., PT08.S2.NMHC., and PT08.S5.O3. as they all have VIF values above 10  
dtm25 <- dtm2[ , c('CO.GT.','NOx.GT.', 'NO2.GT.', 'PT08.S4.NO2.', 'Time', 'PT08.S3.NOx.')]  
names(dtm25)

## [1] "CO.GT." "NOx.GT." "NO2.GT." "PT08.S4.NO2." "Time"   
## [6] "PT08.S3.NOx."

corrplot(cor(dtm25), method = 'square')



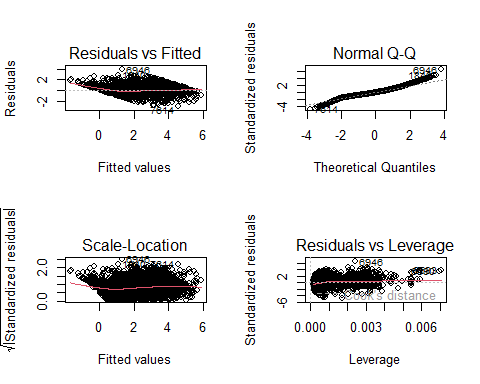
corrplot(cor(dtm25), method = 'number')



m25 <- lm(CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S4.NO2. + Time + PT08.S3.NOx., data = dtm25)  
summary(m25)

##   
## Call:  
## lm(formula = CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S4.NO2. + Time +   
## PT08.S3.NOx., data = dtm25)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8772 -0.3927 -0.0680 0.3264 3.8644   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.330e-01 3.191e-02 29.24 <2e-16 \*\*\*  
## NO2.GT. -2.330e-03 1.302e-04 -17.90 <2e-16 \*\*\*  
## NOx.GT. 3.446e-03 5.587e-05 61.67 <2e-16 \*\*\*  
## PT08.S4.NO2. 1.030e-03 1.593e-05 64.66 <2e-16 \*\*\*  
## Time 2.291e-02 1.092e-03 20.99 <2e-16 \*\*\*  
## PT08.S3.NOx. -1.391e-03 2.637e-05 -52.73 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.6021 on 7426 degrees of freedom  
## Multiple R-squared: 0.7494, Adjusted R-squared: 0.7493   
## F-statistic: 4442 on 5 and 7426 DF, p-value: < 2.2e-16

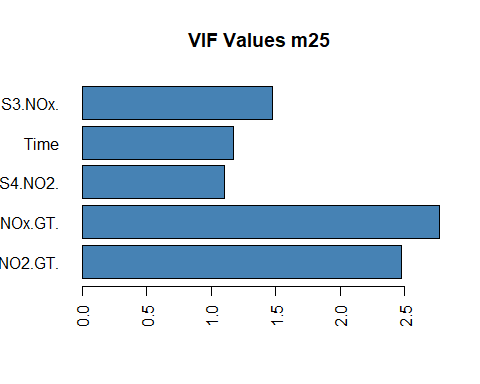
par(mfrow = c(2, 2))  
plot(m25)



vif\_values25 <- vif(m25)  
vif\_values25

## NO2.GT. NOx.GT. PT08.S4.NO2. Time PT08.S3.NOx.   
## 2.478329 2.771912 1.101128 1.173337 1.475125

par(mfrow = c(1,1))  
barplot(vif\_values25, main= "VIF Values m25", horiz = TRUE, col = "steelblue", las=2)  
abline(v = 10, lwd = 3, lty = 2)



# 4. Comment in the R code a justification on why you removed, combined, or left all variables/observations.  
  
# I chose to only keep values which had a high degree of correlation with CO.GT. (abs>0.3) and which had a VIF value under 10.   
# This led to only choosing 'CO.GT.','NOx.GT.', 'NO2.GT.', 'PT08.S4.NO2.', 'Time', and 'PT08.S3.NOx.'  
  
# 5. Kudos bonus if you create dummy variables from character values.  
  
# This step was not needed as the character values (strings) could be converted to numeric as it was a Time variable  
  
# 6. a stepAIC analysis  
  
fit1 <- stepAIC(m1, direction = "both")

## Start: AIC=-12384.44  
## CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH  
##   
## Df Sum of Sq RSS AIC  
## - PT08.S3.NOx. 1 0.021 1398.9 -12386  
## <none> 1398.9 -12384  
## - PT08.S5.O3. 1 2.313 1401.2 -12374  
## - PT08.S2.NMHC. 1 5.027 1403.9 -12360  
## - RH 1 14.070 1412.9 -12312  
## - PT08.S4.NO2. 1 16.759 1415.6 -12298  
## - AH 1 17.508 1416.4 -12294  
## - NMHC.GT. 1 18.931 1417.8 -12286  
## - C6H6.GT. 1 28.210 1427.1 -12238  
## - T 1 28.307 1427.2 -12238  
## - PT08.S1.CO. 1 29.846 1428.7 -12230  
## - Time 1 42.693 1441.6 -12163  
## - NO2.GT. 1 68.994 1467.9 -12029  
## - NOx.GT. 1 228.855 1627.7 -11260  
##   
## Step: AIC=-12386.33  
## CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH +   
## AH  
##   
## Df Sum of Sq RSS AIC  
## <none> 1398.9 -12386  
## + PT08.S3.NOx. 1 0.021 1398.9 -12384  
## - PT08.S5.O3. 1 2.296 1401.2 -12376  
## - PT08.S2.NMHC. 1 7.168 1406.0 -12350  
## - RH 1 14.681 1413.6 -12311  
## - AH 1 17.758 1416.6 -12295  
## - PT08.S4.NO2. 1 18.047 1416.9 -12293  
## - NMHC.GT. 1 19.707 1418.6 -12284  
## - T 1 29.222 1428.1 -12235  
## - PT08.S1.CO. 1 31.828 1430.7 -12221  
## - C6H6.GT. 1 32.492 1431.4 -12218  
## - Time 1 42.702 1441.6 -12165  
## - NO2.GT. 1 69.504 1468.4 -12028  
## - NOx.GT. 1 231.235 1630.1 -11251

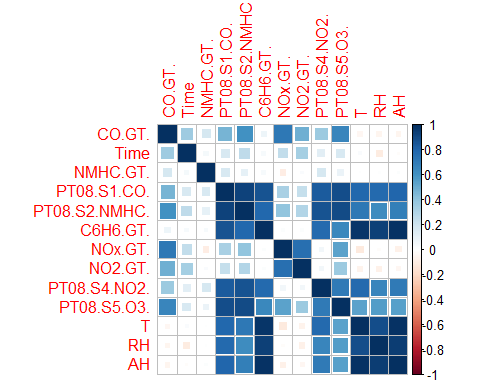
fit1

##   
## Call:  
## lm(formula = CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. +   
## PT08.S2.NMHC. + NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH, data = dtfinal)  
##   
## Coefficients:  
## (Intercept) Time PT08.S1.CO. NMHC.GT. C6H6.GT.   
## -0.7547685 0.0130865 0.0009244 0.0004460 0.0671689   
## PT08.S2.NMHC. NOx.GT. NO2.GT. PT08.S4.NO2. PT08.S5.O3.   
## 0.0009484 0.0020826 -0.0020200 0.0004904 -0.0001314   
## T RH AH   
## -0.0201729 -0.0056031 -0.0568752

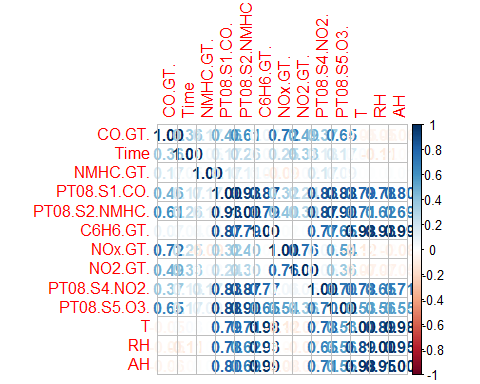
summary(fit1)

##   
## Call:  
## lm(formula = CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. +   
## PT08.S2.NMHC. + NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH, data = dtfinal)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1331 -0.1791 -0.0040 0.1802 3.2209   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.548e-01 8.249e-02 -9.150 < 2e-16 \*\*\*  
## Time 1.309e-02 8.696e-04 15.049 < 2e-16 \*\*\*  
## PT08.S1.CO. 9.243e-04 7.115e-05 12.992 < 2e-16 \*\*\*  
## NMHC.GT. 4.460e-04 4.363e-05 10.223 < 2e-16 \*\*\*  
## C6H6.GT. 6.717e-02 5.117e-03 13.127 < 2e-16 \*\*\*  
## PT08.S2.NMHC. 9.484e-04 1.538e-04 6.166 7.39e-10 \*\*\*  
## NOx.GT. 2.083e-03 5.947e-05 35.019 < 2e-16 \*\*\*  
## NO2.GT. -2.020e-03 1.052e-04 -19.199 < 2e-16 \*\*\*  
## PT08.S4.NO2. 4.904e-04 5.012e-05 9.783 < 2e-16 \*\*\*  
## PT08.S5.O3. -1.314e-04 3.767e-05 -3.490 0.000486 \*\*\*  
## T -2.017e-02 1.620e-03 -12.449 < 2e-16 \*\*\*  
## RH -5.603e-03 6.350e-04 -8.824 < 2e-16 \*\*\*  
## AH -5.688e-02 5.861e-03 -9.705 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4342 on 7419 degrees of freedom  
## Multiple R-squared: 0.8698, Adjusted R-squared: 0.8696   
## F-statistic: 4130 on 12 and 7419 DF, p-value: < 2.2e-16

dtm3 = dtfinal[ , c('CO.GT.', 'Time', 'NMHC.GT.', 'PT08.S1.CO.', 'PT08.S2.NMHC.', 'C6H6.GT.', 'NOx.GT.', 'NO2.GT.', 'PT08.S4.NO2.', 'PT08.S5.O3.', 'T', 'RH', 'AH')]  
corrplot(cor(dtm3), method = 'square')



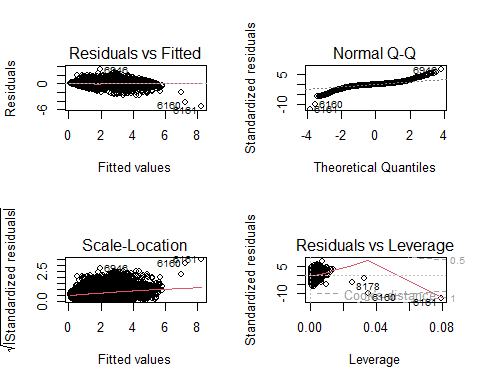
corrplot(cor(dtm3), method = 'number')



m3 <- lm(formula = CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. + NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH + AH, data = dtm3)  
summary(m3)

##   
## Call:  
## lm(formula = CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. +   
## PT08.S2.NMHC. + NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH, data = dtm3)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.1331 -0.1791 -0.0040 0.1802 3.2209   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -7.548e-01 8.249e-02 -9.150 < 2e-16 \*\*\*  
## Time 1.309e-02 8.696e-04 15.049 < 2e-16 \*\*\*  
## PT08.S1.CO. 9.243e-04 7.115e-05 12.992 < 2e-16 \*\*\*  
## NMHC.GT. 4.460e-04 4.363e-05 10.223 < 2e-16 \*\*\*  
## C6H6.GT. 6.717e-02 5.117e-03 13.127 < 2e-16 \*\*\*  
## PT08.S2.NMHC. 9.484e-04 1.538e-04 6.166 7.39e-10 \*\*\*  
## NOx.GT. 2.083e-03 5.947e-05 35.019 < 2e-16 \*\*\*  
## NO2.GT. -2.020e-03 1.052e-04 -19.199 < 2e-16 \*\*\*  
## PT08.S4.NO2. 4.904e-04 5.012e-05 9.783 < 2e-16 \*\*\*  
## PT08.S5.O3. -1.314e-04 3.767e-05 -3.490 0.000486 \*\*\*  
## T -2.017e-02 1.620e-03 -12.449 < 2e-16 \*\*\*  
## RH -5.603e-03 6.350e-04 -8.824 < 2e-16 \*\*\*  
## AH -5.688e-02 5.861e-03 -9.705 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.4342 on 7419 degrees of freedom  
## Multiple R-squared: 0.8698, Adjusted R-squared: 0.8696   
## F-statistic: 4130 on 12 and 7419 DF, p-value: < 2.2e-16

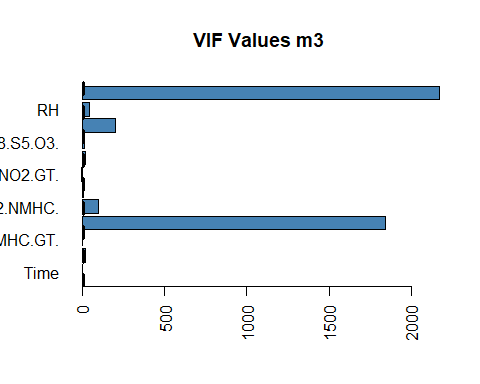
par(mfrow = c(2, 2))  
plot(m3)



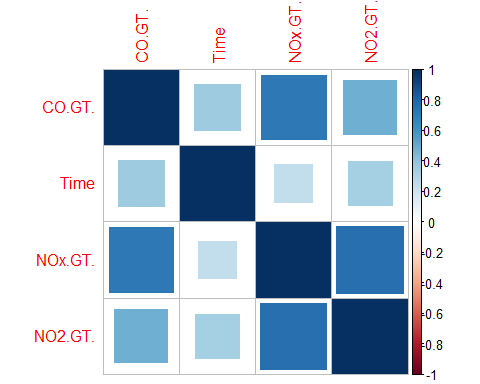
vif\_values3 <- vif(m3)  
vif\_values3

## Time PT08.S1.CO. NMHC.GT. C6H6.GT. PT08.S2.NMHC.   
## 1.431350 20.915550 1.574991 1842.408438 100.150669   
## NOx.GT. NO2.GT. PT08.S4.NO2. PT08.S5.O3. T   
## 6.037808 3.112877 20.962630 10.916972 203.001091   
## RH AH   
## 43.761941 2172.685853

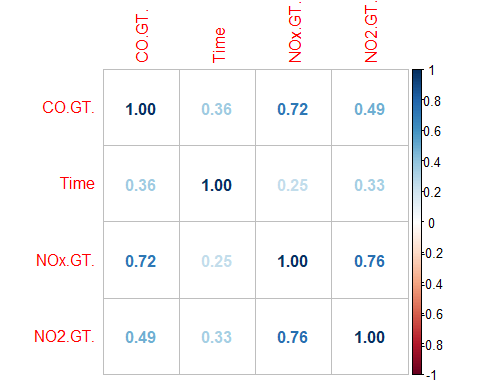
par(mfrow = c(1,1))  
barplot(vif\_values3, main= "VIF Values m3", horiz = TRUE, col = "steelblue", las=2)  
abline(v = 10, lwd = 3, lty = 2)



# based on the correlation values and VIF values, I will drop from dtm3 to get dtm4 model  
  
dtm4 = dtm3[ , c('CO.GT.', 'Time', 'NOx.GT.', 'NO2.GT.')]  
corrplot(cor(dtm4), method = 'square')



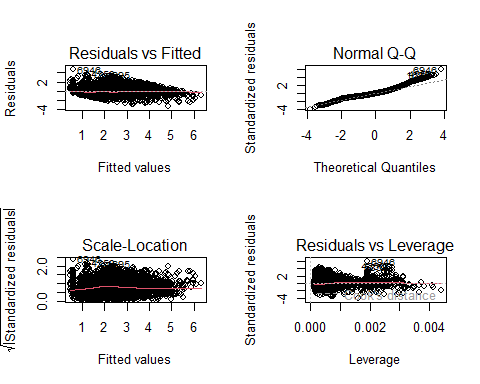
corrplot(cor(dtm4), method = 'number')



m4 <- lm(formula = CO.GT. ~ Time + NOx.GT. + NO2.GT., data = dtm4)  
summary(m4)

##   
## Call:  
## lm(formula = CO.GT. ~ Time + NOx.GT. + NO2.GT., data = dtm4)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.1713 -0.5388 -0.1249 0.3964 4.6214   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.243e-01 1.896e-02 43.48 <2e-16 \*\*\*  
## Time 3.914e-02 1.410e-03 27.75 <2e-16 \*\*\*  
## NOx.GT. 4.696e-03 6.766e-05 69.41 <2e-16 \*\*\*  
## NO2.GT. -2.881e-03 1.714e-04 -16.81 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7944 on 7428 degrees of freedom  
## Multiple R-squared: 0.5638, Adjusted R-squared: 0.5636   
## F-statistic: 3200 on 3 and 7428 DF, p-value: < 2.2e-16

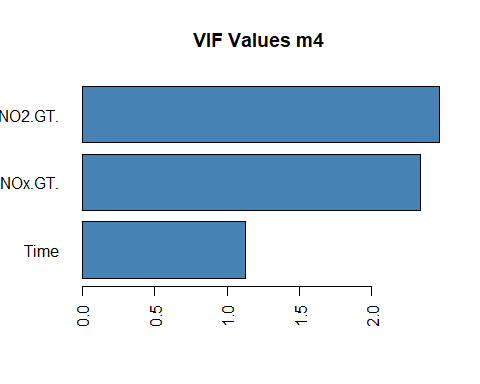
par(mfrow = c(2, 2))  
plot(m4)



vif\_values4 <- vif(m4)  
vif\_values4

## Time NOx.GT. NO2.GT.   
## 1.124989 2.335520 2.467879

par(mfrow = c(1,1))  
barplot(vif\_values4, main= "VIF Values m4", horiz = TRUE, col = "steelblue", las=2)  
abline(v = 10, lwd = 3, lty = 2)



# 7. Summary, residual plots, and a VIF analysis of the final best model you created.  
  
anova(m1, m2)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH  
## Model 2: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S5.O3. + PT08.S2.NMHC. + PT08.S1.CO. +   
## PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7418 1398.9   
## 2 7423 2480.0 -5 -1081.1 1146.6 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m1, m25)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH  
## Model 2: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7418 1398.9   
## 2 7426 2692.1 -8 -1293.2 857.24 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m1, m3)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH  
## Model 2: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH +   
## AH  
## Res.Df RSS Df Sum of Sq F Pr(>F)  
## 1 7418 1398.9   
## 2 7419 1398.9 -1 -0.021356 0.1133 0.7365

anova(m1, m4)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + PT08.S3.NOx. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. +   
## T + RH + AH  
## Model 2: CO.GT. ~ Time + NOx.GT. + NO2.GT.  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7418 1398.9   
## 2 7428 4687.2 -10 -3288.4 1743.8 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m2, m25)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S5.O3. + PT08.S2.NMHC. + PT08.S1.CO. +   
## PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Model 2: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7423 2480.0   
## 2 7426 2692.1 -3 -212.11 211.63 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m2, m3)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S5.O3. + PT08.S2.NMHC. + PT08.S1.CO. +   
## PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Model 2: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH +   
## AH  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7423 2480.0   
## 2 7419 1398.9 4 1081.1 1433.4 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m2, m4)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S5.O3. + PT08.S2.NMHC. + PT08.S1.CO. +   
## PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Model 2: CO.GT. ~ Time + NOx.GT. + NO2.GT.  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7423 2480.0   
## 2 7428 4687.2 -5 -2207.2 1321.3 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m25, m3)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Model 2: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH +   
## AH  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7426 2692.1   
## 2 7419 1398.9 7 1293.2 979.8 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m25, m4)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ NO2.GT. + NOx.GT. + PT08.S4.NO2. + Time + PT08.S3.NOx.  
## Model 2: CO.GT. ~ Time + NOx.GT. + NO2.GT.  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7426 2692.1   
## 2 7428 4687.2 -2 -1995.1 2751.7 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

anova(m3, m4)

## Analysis of Variance Table  
##   
## Model 1: CO.GT. ~ Time + PT08.S1.CO. + NMHC.GT. + C6H6.GT. + PT08.S2.NMHC. +   
## NOx.GT. + NO2.GT. + PT08.S4.NO2. + PT08.S5.O3. + T + RH +   
## AH  
## Model 2: CO.GT. ~ Time + NOx.GT. + NO2.GT.  
## Res.Df RSS Df Sum of Sq F Pr(>F)   
## 1 7419 1398.9   
## 2 7428 4687.2 -9 -3288.4 1937.8 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# 8. AIC scores of all models built

#### Conclusion

1. What general conclusions can you make from the results?
2. What does this mean?
3. How is this significant?
4. How did the testing go?
5. What shoud be done different/going forward?