

Ozone Final Models

2023-12-08

```
#Imports

library(ISLR)
library(dplyr)

## 
## Attaching package: 'dplyr'

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

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

library(MASS)

## 
## Attaching package: 'MASS'

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

library(car) # for VIF

## Loading required package: carData

## 
## Attaching package: 'car'

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

library(corrplot) #correlation plot

## corrplot 0.92 loaded
```

```

library(RColorBrewer) #for corrplot
library(caret) #caret cross validation machine learning

## Loading required package: ggplot2

## Loading required package: lattice

library(leaps) #linear feature selection
library(glmnet) #for LASSO regression

## Loading required package: Matrix

## Loaded glmnet 4.1-8

library(mgcv) #GAM

## Loading required package: nlme

##
## Attaching package: 'nlme'

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

## This is mgcv 1.9-0. For overview type 'help("mgcv-package")'.

library(randomForest) #random forest

## randomForest 4.7-1.1

## Type rfNews() to see new features/changes/bug fixes.

##
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':
##      margin

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

library(nnet) # for ANN

##
## Attaching package: 'nnet'

## The following object is masked from 'package:mgcv':
##      multinom

```

```

library(gbm) #for boosted trees

## Loaded gbm 2.1.8.1

library(ggformula) #for plots

## Loading required package: scales

## Loading required package: ggridges

##
## New to ggformula? Try the tutorials:
##   learnr::run_tutorial("introduction", package = "ggformula")
##   learnr::run_tutorial("refining", package = "ggformula")

library(DMwR2) #imputation

## Registered S3 method overwritten by 'quantmod':
##   method           from
##   as.zoo.data.frame zoo

library(naniar) #missing values
library(VIM) #for aggr()

## Loading required package: colorspace

## Loading required package: grid

## VIM is ready to use.

## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues

##
## Attaching package: 'VIM'

## The following object is masked from 'package:DMwR2':
##   kNN

## The following object is masked from 'package:datasets':
##   sleep

```

Load & Re-organize Data

```

setwd('/Users/paultrygstad/Documents/Grad School/Classes/Fall 2023/DS785/Data/Data Cleaning/Cleaned Data')
data <- read.csv("Cleaned_Model_Data_final.csv")
head(data)

```

```

##      AFA_Ozone_1.HOUR_Arithmetic.Mean
## 1          0.024000
## 2          0.030417
## 3          0.022792
## 4          0.012609
## 5          0.030000
## 6          0.022833
##      AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean
## 1          0.030588
## 2          0.027059
## 3          0.019588
## 4          0.013824
## 5          0.033294
## 6          0.018118
##      HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean
## 1          2.020833
## 2          1.745833
## 3          2.041667
## 4          2.450000
## 5          2.716667
## 6          2.666667
##      HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean
## 1          242.7500
## 2          222.9167
## 3          258.8333
## 4          257.7917
## 5          239.5833
## 6          237.7917
##      HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean
## 1          0.542105
## 2          0.720833
## 3          0.795833
## 4          0.825000
## 5          0.733333
## 6          0.837500
##      HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean
## 1          0.587917
## 2          0.790000
## 3          0.742917
## 4          0.805217
## 5          0.732083
## 6          0.829583
##      HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## 1          245.0000
## 2          225.3750
## 3          258.7083
## 4          257.5000
## 5          226.5417
## 6          252.7500
##      HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean
## 1          1.641667
## 2          1.379167
## 3          1.866667
## 4          2.250000

```

```

## 5 2.237500
## 6 2.191667
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean
## 1 2.221739
## 2 1.756522
## 3 4.660870
## 4 5.645455
## 5 3.339130
## 6 5.860870
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean
## 1 2.385714
## 2 1.771429
## 3 4.900000
## 4 6.450000
## 5 3.357143
## 6 6.300000
## HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean
## 1 5.656522
## 2 2.495652
## 3 8.278261
## 4 10.059091
## 5 5.660870
## 6 11.156522
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
## 1 50.41667
## 2 35.45833
## 3 41.83333
## 4 50.33333
## 5 43.83333
## 6 53.87500
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
## 1 21.12500
## 2 27.41667
## 3 29.04167
## 4 29.25000
## 5 35.79167
## 6 34.04167
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
## 1 37.68750
## 2 40.37083
## 3 25.51250
## 4 23.03333
## 5 37.73750
## 6 37.76250
## MAN_Ozone_1.HOUR_Arithmetic.Mean
## 1 0.037333
## 2 0.043667
## 3 0.035417
## 4 0.027783
## 5 0.035417
## 6 0.033792
## MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean Average.Mixing.Height
## 1 0.041471 179
## 2 0.040529 518

```

```

## 3                               0.033059      NA
## 4                               0.025706      NA
## 5                               0.036588      NA
## 6                               0.030235      NA
##   Met_Lat  Met_long Met_elevation Average.daily.wind.speed Precipitation
## 1 38.80949 -104.6887     1884.2          5.14        0
## 2 38.80949 -104.6887     1884.2          5.14        0
## 3 38.80949 -104.6887     1884.2          4.70        0
## 4 38.80949 -104.6887     1884.2          6.26        0
## 5 38.80949 -104.6887     1884.2          8.05        0
## 6 38.80949 -104.6887     1884.2          6.49        0
##   Snowfall Snow.depth Temperature.Average Direction.of.fastest.2.minute.wind
## 1       0         0           19            20
## 2       0         0           27           110
## 3       0         0           29            20
## 4       0         0           29           170
## 5       0         0           35            20
## 6       0         0           36           150
##   Direction.of.fastest.5.second.wind Fastest.2.minute.wind.speed
## 1                           20            10.1
## 2                           110           13.0
## 3                           10            10.1
## 4                           190           10.1
## 5                           340           19.9
## 6                           150           18.1
##   Fastest.5.second.wind.speed      Date
## 1                         12.1 2016-01-01
## 2                         14.1 2016-01-02
## 3                         12.1 2016-01-03
## 4                         14.1 2016-01-04
## 5                         23.9 2016-01-05
## 6                         21.0 2016-01-06

```

```
colnames(data)
```

```

## [1] "AFA_Ozone_1.HOUR_Arithmetic.Mean"
## [2] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [3] "HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean"
## [4] "HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean"
## [5] "HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean"
## [6] "HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean"
## [7] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean"
## [8] "HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean"
## [9] "HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean"
## [10] "HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean"
## [11] "HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean"
## [12] "HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean"
## [13] "HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean"
## [14] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean"
## [15] "MAN_Ozone_1.HOUR_Arithmetic.Mean"
## [16] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [17] "Average.Mixing.Height"
## [18] "Met_Lat"
## [19] "Met_long"

```

```

## [20] "Met_elevation"
## [21] "Average.daily.wind.speed"
## [22] "Precipitation"
## [23] "Snowfall"
## [24] "Snow.depth"
## [25] "Temperature.Average"
## [26] "Direction.of.fastest.2.minute.wind"
## [27] "Direction.of.fastest.5.second.wind"
## [28] "Fastest.2.minute.wind.speed"
## [29] "Fastest.5.second.wind.speed"
## [30] "Date"

#Missing Values

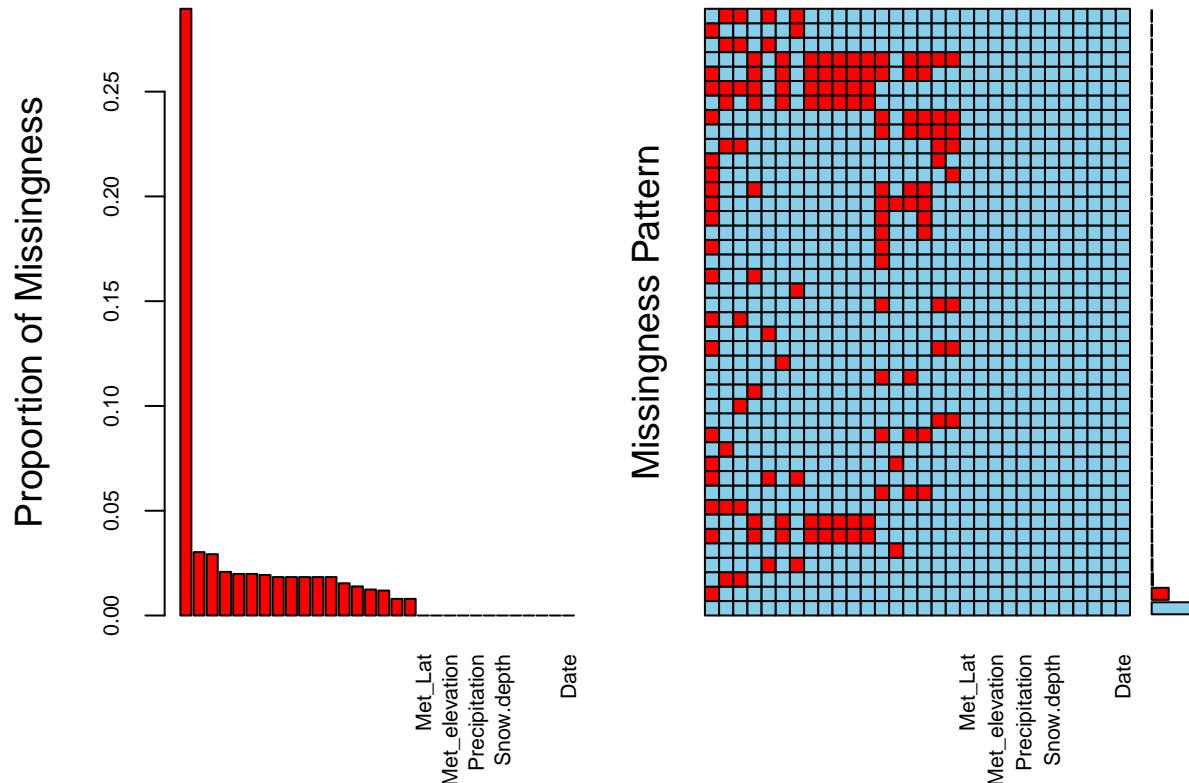
data$date = as.Date(data$date)
aggr(data, numbers=TRUE, sortVars=TRUE, cex.axis=.7, gap=3,
      ylab=c("Proportion of Missingness", "Missingness Pattern"))

```

```

## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)

```



```

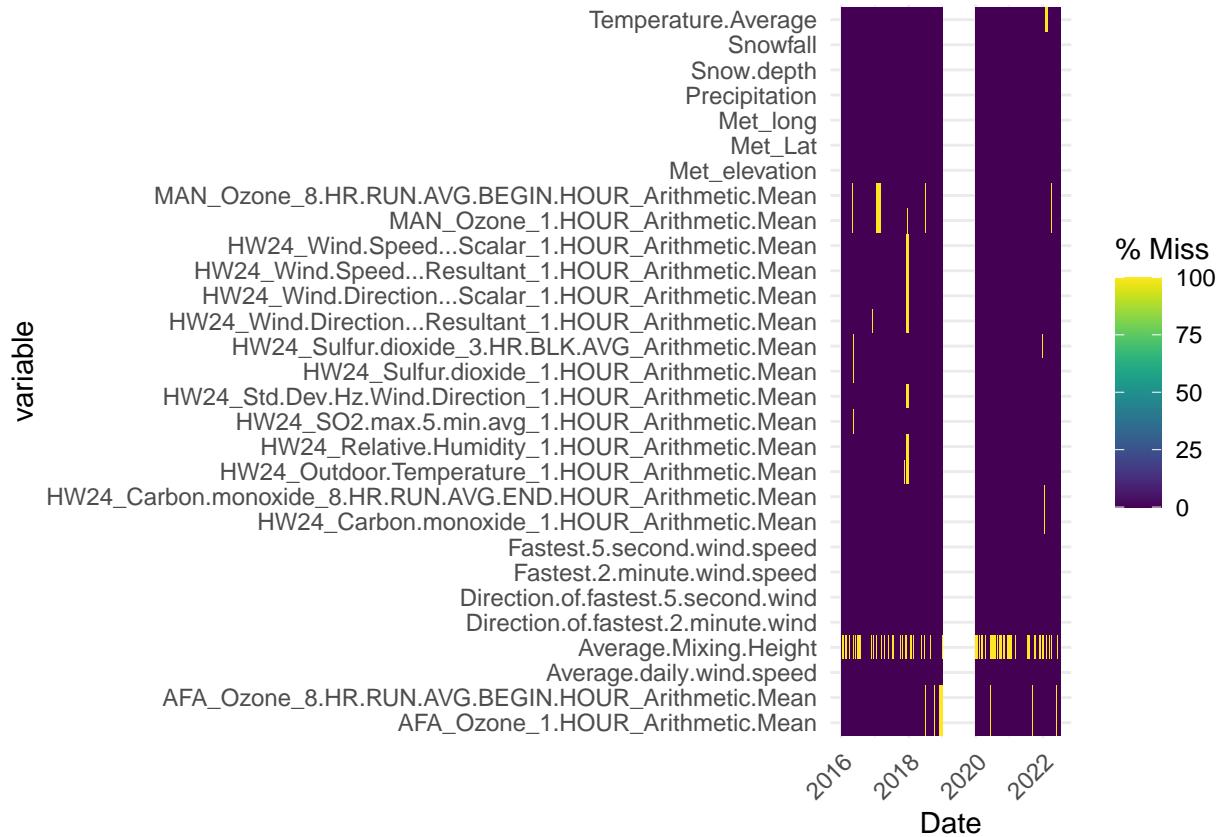
##
## Variables sorted by number of missings:
##
```

Variable	Count
Met_Lat	1
Met_elevation	1
Precipitation	1
Snow.depth	1
Date	1

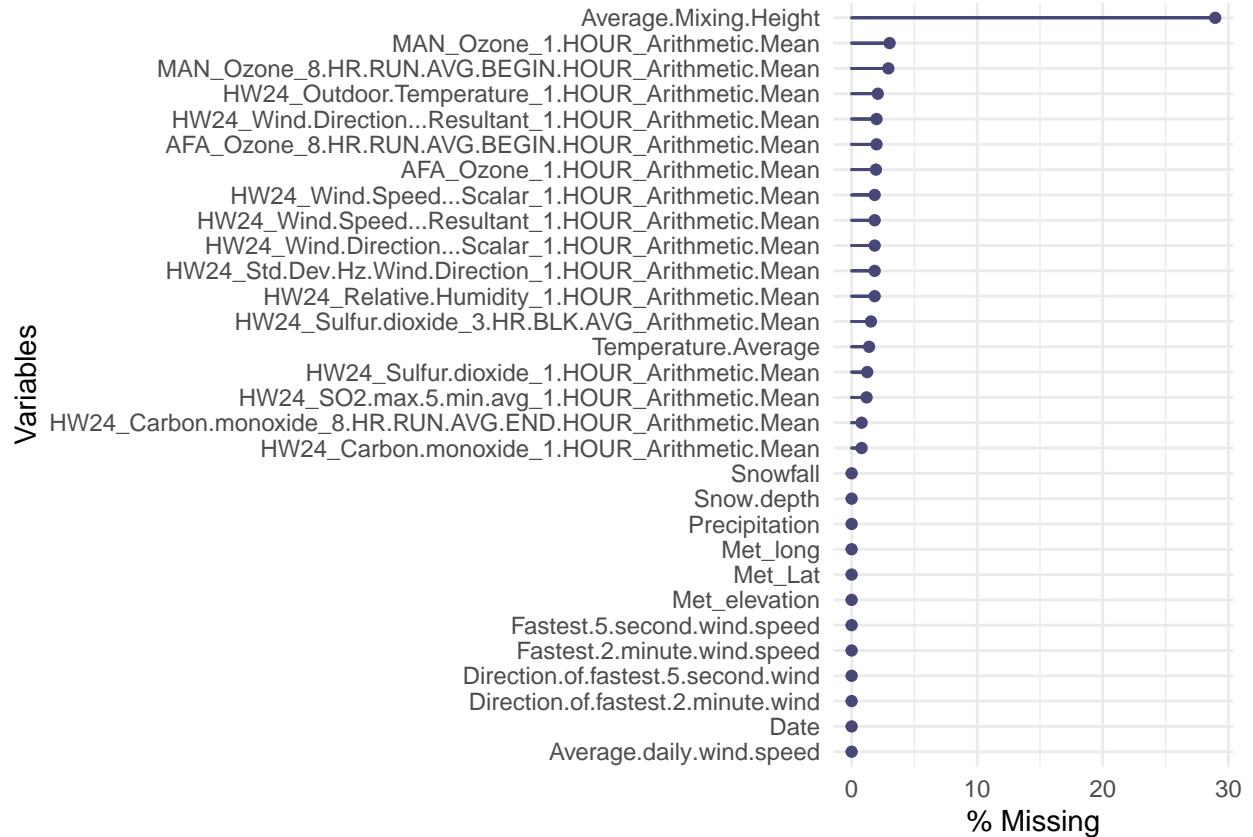
```

##                               Average.Mixing.Height 0.289538919
##                               MAN_Ozone_1.HOUR_Arithmetic.Mean 0.030242935
##                               MAN_Ozone_8.HR.RUN.AVG.BEGIN.HOUR_Arithmetic.Mean 0.029251363
##                               HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean 0.020823004
##                               AFA_Ozone_8.HR.RUN.AVG.BEGIN.HOUR_Arithmetic.Mean 0.019831433
##                               HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 0.019831433
##                               AFA_Ozone_1.HOUR_Arithmetic.Mean 0.019335647
##                               HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean 0.018344075
##                               HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean 0.018344075
##                               HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean 0.018344075
##                               HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean 0.018344075
##                               HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean 0.018344075
##                               HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetric.Mean 0.015369360
##                               Temperature.Average 0.013882003
##                               HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean 0.012394646
##                               HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean 0.011898860
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 0.007932573
##                               HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean 0.007932573
##                               Met_Lat 0.0000000000
##                               Met_long 0.0000000000
##                               Met_elevation 0.0000000000
##                               Average.daily.wind.speed 0.0000000000
##                               Precipitation 0.0000000000
##                               Snowfall 0.0000000000
##                               Snow.depth 0.0000000000
##                               Direction.of.fastest.2.minute.wind 0.0000000000
##                               Direction.of.fastest.5.second.wind 0.0000000000
##                               Fastest.2.minute.wind.speed 0.0000000000
##                               Fastest.5.second.wind.speed 0.0000000000
##                               Date 0.0000000000
##
```

```
gg_miss_fct(data, Date)
```



```
gg_miss_var(data, show_pct = TRUE)
```



```
#KNN Imputation
```

```
# KNN Imputation
set.seed(28) # for reproducibility
data_impute = knnImputation(data[,c(-30)], k=2, scale=TRUE) #use 10 as default
head(data_impute)
```

```
##   AFA_Ozone_1.HOUR_Arithmetic.Mean
## 1          0.024000
## 2          0.030417
## 3          0.022792
## 4          0.012609
## 5          0.030000
## 6          0.022833
##   AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean
## 1          0.030588
## 2          0.027059
## 3          0.019588
## 4          0.013824
## 5          0.033294
## 6          0.018118
##   HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean
## 1          2.020833
## 2          1.745833
## 3          2.041667
## 4          2.450000
```

```

## 5 2.716667
## 6 2.666667
## HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean
## 1 242.7500
## 2 222.9167
## 3 258.8333
## 4 257.7917
## 5 239.5833
## 6 237.7917
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean
## 1 0.542105
## 2 0.720833
## 3 0.795833
## 4 0.825000
## 5 0.733333
## 6 0.837500
## HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean
## 1 0.587917
## 2 0.790000
## 3 0.742917
## 4 0.805217
## 5 0.732083
## 6 0.829583
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## 1 245.0000
## 2 225.3750
## 3 258.7083
## 4 257.5000
## 5 226.5417
## 6 252.7500
## HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean
## 1 1.641667
## 2 1.379167
## 3 1.866667
## 4 2.250000
## 5 2.237500
## 6 2.191667
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean
## 1 2.221739
## 2 1.756522
## 3 4.660870
## 4 5.645455
## 5 3.339130
## 6 5.860870
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean
## 1 2.385714
## 2 1.771429
## 3 4.900000
## 4 6.450000
## 5 3.357143
## 6 6.300000
## HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean
## 1 5.656522
## 2 2.495652

```

```

## 3 8.278261
## 4 10.059091
## 5 5.660870
## 6 11.156522
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
## 1 50.41667
## 2 35.45833
## 3 41.83333
## 4 50.33333
## 5 43.83333
## 6 53.87500
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
## 1 21.12500
## 2 27.41667
## 3 29.04167
## 4 29.25000
## 5 35.79167
## 6 34.04167
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
## 1 37.68750
## 2 40.37083
## 3 25.51250
## 4 23.03333
## 5 37.73750
## 6 37.76250
## MAN_Ozone_1.HOUR_Arithmetic.Mean
## 1 0.037333
## 2 0.043667
## 3 0.035417
## 4 0.027783
## 5 0.035417
## 6 0.033792
## MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean Average.Mixing.Height
## 1 0.041471 179.0000
## 2 0.040529 518.0000
## 3 0.033059 595.4635
## 4 0.025706 952.7986
## 5 0.036588 1053.1539
## 6 0.030235 1008.0490
## Met Lat Met_long Met_elevation Average.daily.wind.speed Precipitation
## 1 38.80949 -104.6887 1884.2 5.14 0
## 2 38.80949 -104.6887 1884.2 5.14 0
## 3 38.80949 -104.6887 1884.2 4.70 0
## 4 38.80949 -104.6887 1884.2 6.26 0
## 5 38.80949 -104.6887 1884.2 8.05 0
## 6 38.80949 -104.6887 1884.2 6.49 0
## Snowfall Snow.depth Temperature.Average Direction.of.fastest.2.minute.wind
## 1 0 0 19 20
## 2 0 0 27 110
## 3 0 0 29 20
## 4 0 0 29 170
## 5 0 0 35 20
## 6 0 0 36 150
## Direction.of.fastest.5.second.wind Fastest.2.minute.wind.speed

```

```

## 1                      20                  10.1
## 2                     110                 13.0
## 3                      10                  10.1
## 4                     190                 10.1
## 5                     340                 19.9
## 6                     150                 18.1
##   Fastest.5.second.wind.speed
## 1                      12.1
## 2                     14.1
## 3                     12.1
## 4                     14.1
## 5                     23.9
## 6                     21.0

data_impute["Date"] = data["Date"] #add date back into dataframe
#plot imputed values
scatter_logic = is.na(data$Average.Mixing.Height)
scatter_data = data
data_impute$Average.Mixing.Height[scatter_logic]

## [1] 595.46347 952.79857 1053.15391 1008.04902 244.16675 374.55435
## [7] 456.29731 780.81173 175.72367 705.31556 830.10040 173.26005
## [13] 560.81833 70.79357 817.66624 1056.83708 289.95932 264.95282
## [19] 308.67799 357.24690 101.92164 1001.41289 465.71517 1807.42269
## [25] 1358.24490 2297.66510 729.03260 472.29611 1738.50932 143.29573
## [31] 2329.72919 89.01166 1447.20473 2330.05058 2974.38549 1746.01506
## [37] 1268.92975 1296.88655 2476.85010 3192.24727 1649.57551 1508.79073
## [43] 1953.57113 512.60831 421.64243 1571.98723 2210.62410 3048.37713
## [49] 2272.10653 1330.00172 3838.31135 3020.07682 1938.36317 3638.48455
## [55] 3119.86864 1077.14523 1551.22275 3487.14578 1400.19997 832.45717
## [61] 2754.71879 1702.15146 2116.23063 1518.79033 1142.17374 1790.80808
## [67] 2428.54352 967.64042 3461.89087 1387.21789 1457.44306 2003.45308
## [73] 961.97548 3056.32711 3262.24665 2431.25737 1125.24380 2199.12714
## [79] 1761.79811 3987.20840 2825.13753 3547.37541 2449.98979 1357.15809
## [85] 2304.50996 2079.19652 3366.08390 1280.37730 1893.36072 425.63805
## [91] 2169.62564 1456.27440 576.07031 1240.94022 1916.77787 1482.61731
## [97] 793.74974 939.87723 575.72378 278.62776 808.14445 684.69976
## [103] 25.73033 838.17895 542.15073 227.27706 60.09570 398.64991
## [109] 257.71515 482.98481 591.59360 289.52572 1795.32856 419.36737
## [115] 963.30287 274.29398 1451.00162 442.96163 1047.46095 2174.07800
## [121] 2112.84312 673.97754 430.87700 1550.19253 2343.69086 1219.94000
## [127] 1815.21849 3196.88058 1863.05413 3120.60331 2072.59724 3123.52232
## [133] 2886.35202 3296.33823 2075.04747 1679.38852 2732.28628 762.52689
## [139] 3632.45093 2009.49690 3924.60103 2174.16741 3713.71281 3813.20194
## [145] 2683.86252 2244.74068 407.33173 2641.92325 1324.35696 1501.77632
## [151] 2277.11312 2615.37024 2418.29418 1316.70249 2416.08589 2015.06536
## [157] 2856.39128 387.19078 1351.17056 1883.47032 1407.15221 1536.79862
## [163] 1267.80831 726.24288 1282.54875 576.35950 721.20522 759.64002
## [169] 528.72457 508.65070 2625.10147 130.31302 820.98144 1086.61940
## [175] 796.66134 476.39081 848.48929 491.73547 2622.01951 1843.04253
## [181] 484.60532 233.96958 224.34970 526.58203 1824.70522 78.59775
## [187] 386.54075 429.87925 1973.32638 417.29971 273.67194 1036.26730
## [193] 1392.92010 555.71347 395.73725 1023.06103 746.41984 303.59080
## [199] 512.88424 255.76200 1320.18789 135.69955 1028.79535 1071.02350

```

```

## [205] 1337.11991 439.40505 197.88751 3731.97589 174.10801 1045.43311
## [211] 820.99246 926.90980 1508.11255 286.89464 909.81172 1145.95849
## [217] 242.69737 2252.61928 429.24495 1478.92760 1403.48614 412.70877
## [223] 1364.68914 1697.13039 468.00296 2552.94190 496.45583 2158.28149
## [229] 2099.05238 2905.82252 2105.96256 3657.59792 1970.81088 1943.50723
## [235] 2929.85066 2806.44793 1997.19993 2462.86689 3085.40724 3883.75471
## [241] 1987.94534 3790.74441 3423.22935 2295.98104 4090.16919 1836.25803
## [247] 2358.94778 1747.83787 3038.84453 1526.63361 3173.77992 3511.30456
## [253] 560.01138 3580.38707 1931.24227 3653.83380 534.27856 604.34528
## [259] 987.69897 999.18433 880.69830 902.13687 707.75988 527.80835
## [265] 1882.75802 615.95306 683.09647 951.06164 1114.62994 637.36014
## [271] 1021.33041 2204.59752 1776.90470 949.55748 770.50176 1394.87773
## [277] 1693.45743 1102.07219 1968.76142 1096.06763 520.42701 848.98621
## [283] 319.20983 764.83237 641.80249 1174.58006 576.77335 1097.42897
## [289] 1259.65666 988.04587 683.69718 1873.01619 578.83634 3471.56836
## [295] 610.59420 538.30600 230.14720 1360.65807 1131.93162 245.50639
## [301] 1447.33150 2353.28112 758.01968 2277.73820 2281.24981 2572.40682
## [307] 1920.46333 732.10406 812.44345 1588.27245 601.89891 2410.74080
## [313] 1239.26923 3474.79791 2277.54562 2639.99741 1543.03938 2658.55959
## [319] 2828.91266 2455.99352 4112.51840 1632.83693 939.64128 2830.64501
## [325] 2005.85532 3125.45336 3158.05355 3392.24156 3364.21208 2762.49565
## [331] 3142.24773 2739.13391 2929.45836 2544.56694 2610.41933 3540.10999
## [337] 1465.05073 1878.51242 3615.05447 2653.23429 1600.19141 2923.45879
## [343] 4019.17217 2321.45112 2230.25105 2752.50488 1809.15853 1774.86238
## [349] 2207.62863 1013.28963 1400.26461 2960.15114 2157.40331 496.89908
## [355] 293.48490 2587.89486 1491.97611 2299.10387 1719.38549 1116.15721
## [361] 2111.87292 629.18313 2521.92815 1622.68717 2729.48798 3608.93122
## [367] 795.47402 1211.34621 2231.59563 1223.91485 417.31489 313.55072
## [373] 1190.40799 356.19895 953.42131 2432.89495 964.68184 387.79296
## [379] 390.67882 584.95119 998.27700 636.94995 1322.93439 989.44819
## [385] 938.31085 793.53545 445.60795 309.68560 688.09184 431.41259
## [391] 537.22034 452.69409 422.88316 877.92826 1776.09353 1407.00691
## [397] 1892.34030 1095.98864 711.05807 1352.10787 427.70168 564.73260
## [403] 941.06036 657.94003 410.76242 2809.74273 928.68595 748.97972
## [409] 405.31887 476.02809 1981.13938 541.36216 795.71750 148.04798
## [415] 1083.25671 1188.72018 921.37088 523.95180 695.23356 452.38369
## [421] 2240.00262 3097.93606 870.17998 1037.96521 950.45868 2064.33789
## [427] 2645.53868 971.19325 869.39579 2471.43283 2216.30195 616.07230
## [433] 756.02077 2107.79082 550.33443 1634.70561 2715.69247 2291.87410
## [439] 1239.66839 1067.60529 1365.68907 2866.93063 2746.91416 1049.17853
## [445] 1546.27995 1777.71832 2460.53591 1891.69857 1912.32979 1323.13551
## [451] 1386.93668 2833.64139 2026.96458 2050.58290 2092.22602 2846.90404
## [457] 2676.57328 1937.02489 949.27830 148.07992 166.29875 2513.94312
## [463] 4046.86584 942.44692 2747.98735 1210.07133 2138.12005 1703.59954
## [469] 1703.70425 3087.86932 2501.90274 163.34180 2659.67075 1253.41919
## [475] 1996.61349 1971.43024 1750.45603 1989.12044 2542.55231 1555.56476
## [481] 1175.96246 1745.94161 1089.62731 3129.10329 2876.56572 2470.38302
## [487] 824.42822 3813.11074 697.19745 1338.84701 2195.52016 2982.28173
## [493] 2283.35901 354.79481 2458.97004 2368.98569 2027.44232 2288.33317
## [499] 1435.54880 1944.52960 931.70407 2104.21139 2251.34290 2113.07522
## [505] 3830.56006 348.40302 2256.15898 255.42188 1900.84242 1343.54361
## [511] 1348.47314 2145.01346 721.37140 1135.22207 779.64223 519.19153
## [517] 514.13551 1297.82816 220.70931 433.92070 931.34565 2314.17754
## [523] 1352.61158 540.63481 148.73434 2249.01673 560.61584 249.46170

```

```

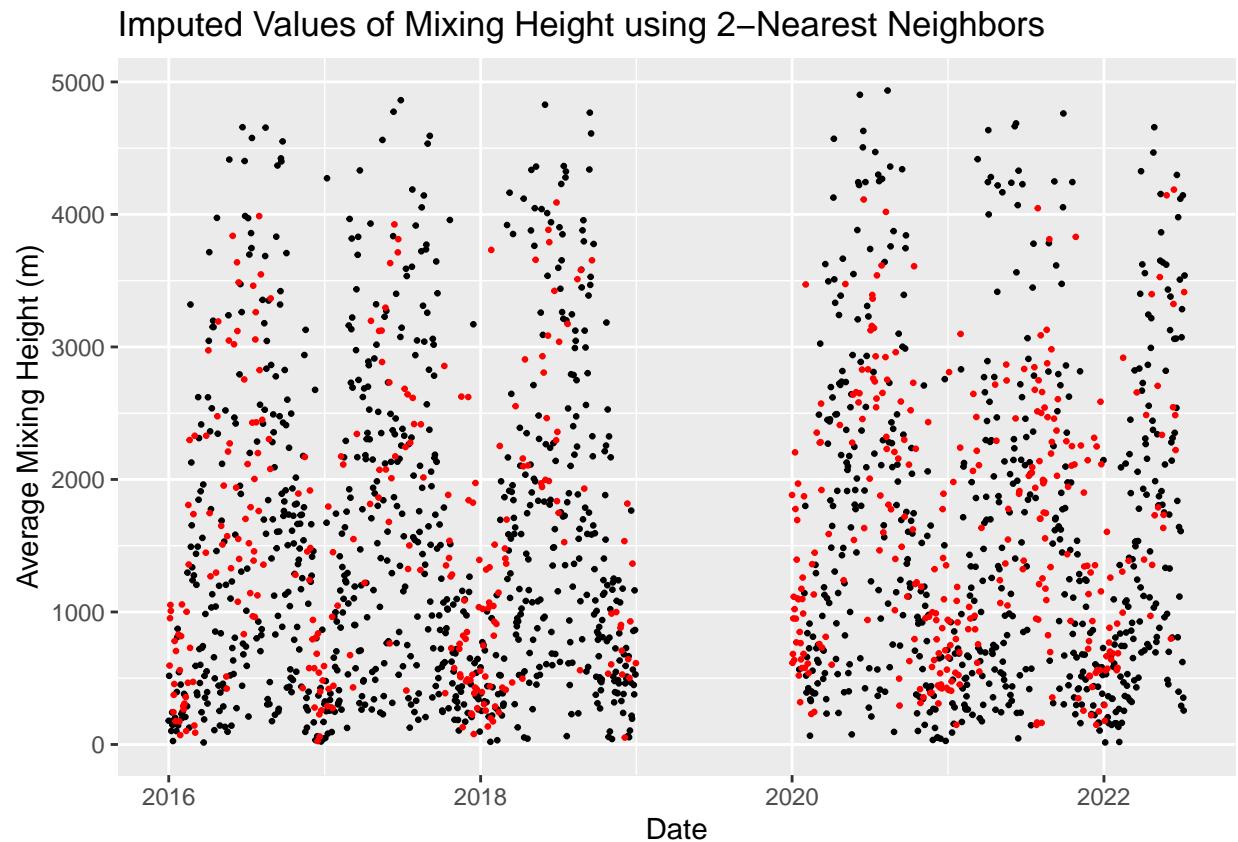
## [529] 561.08669 2586.39402 1155.18213 2114.97040 296.17007 178.52258
## [535] 558.55062 666.01939 1604.44269 571.57313 769.43710 880.62317
## [541] 701.33220 1259.11809 640.82238 686.50867 560.25034 912.37167
## [547] 578.31408 688.24078 1304.29462 994.96895 2918.37340 1386.62442
## [553] 2656.73528 1395.25683 2486.28223 970.49294 3398.53492 1355.08853
## [559] 1729.63648 2706.31061 1789.41090 3527.25426 2335.97525 1634.30883
## [565] 1732.67320 4144.36210 798.64073 2545.21180 3324.42127 4187.14374
## [571] 2485.14939 2221.86857 3413.99887 2605.53878 2581.53679 2650.33565
## [577] 1534.68848 52.15289 1817.96742 643.62624 502.27573 928.87202
## [583] 1364.88026 613.70701

```

```

scatter_data["Average.Mixing.Height.Imputed"] = scatter_data$Average.Mixing.Height
for (ii in 1:length(scatter_logic)) {
  if (scatter_logic[ii] == FALSE) {
    scatter_data$Average.Mixing.Height.Imputed[ii] = NA
  } else {
    scatter_data$Average.Mixing.Height.Imputed[ii] = data_impute$Average.Mixing.Height[ii]
  }
}
#plot
ggplot(scatter_data, aes(y=Average.Mixing.Height, x=Date)) + geom_point(size=.5) + geom_point(aes(y=Ave

```



```
#Reorganize and filter Data
```

```

#move ozone measurement [16]
#drop latitude [18], longitude [19], elevation [20]

```

```

#drop 1 hour ozone for AFA [1] and MAN [15]
colnames(data_impute)

## [1] "AFA_Ozone_1.HOUR_Arithmetic.Mean"
## [2] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [3] "HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean"
## [4] "HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean"
## [5] "HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean"
## [6] "HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean"
## [7] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean"
## [8] "HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean"
## [9] "HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean"
## [10] "HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetric.Mean"
## [11] "HW24_SO2.max.5.min.avg_1.HOUR_Arithmetric.Mean"
## [12] "HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean"
## [13] "HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean"
## [14] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean"
## [15] "MAN_Ozone_1.HOUR_Arithmetric.Mean"
## [16] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetric.Mean"
## [17] "Average.Mixing.Height"
## [18] "Met_Lat"
## [19] "Met_long"
## [20] "Met_elevation"
## [21] "Average.daily.wind.speed"
## [22] "Precipitation"
## [23] "Snowfall"
## [24] "Snow.depth"
## [25] "Temperature.Average"
## [26] "Direction.of.fastest.2.minute.wind"
## [27] "Direction.of.fastest.5.second.wind"
## [28] "Fastest.2.minute.wind.speed"
## [29] "Fastest.5.second.wind.speed"
## [30] "Date"

data_filtered = data_impute[,c(2,16,3,4,5,6,7,8,9,10,11,12,13,14,17,21,22,23,24,25,26,27,28,29,30)]
#columns
colnames(data_filtered)

## [1] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [2] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [3] "HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean"
## [4] "HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean"
## [5] "HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean"
## [6] "HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean"
## [7] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean"
## [8] "HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean"
## [9] "HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean"
## [10] "HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetric.Mean"
## [11] "HW24_SO2.max.5.min.avg_1.HOUR_Arithmetric.Mean"
## [12] "HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean"
## [13] "HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean"
## [14] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean"
```

```

## [15] "Average.Mixing.Height"
## [16] "Average.daily.wind.speed"
## [17] "Precipitation"
## [18] "Snowfall"
## [19] "Snow.depth"
## [20] "Temperature.Average"
## [21] "Direction.of.fastest.2.minute.wind"
## [22] "Direction.of.fastest.5.second.wind"
## [23] "Fastest.2.minute.wind.speed"
## [24] "Fastest.5.second.wind.speed"
## [25] "Date"

#data_filtered$Date = as.Date(data_filtered$Date) # change Date column data type

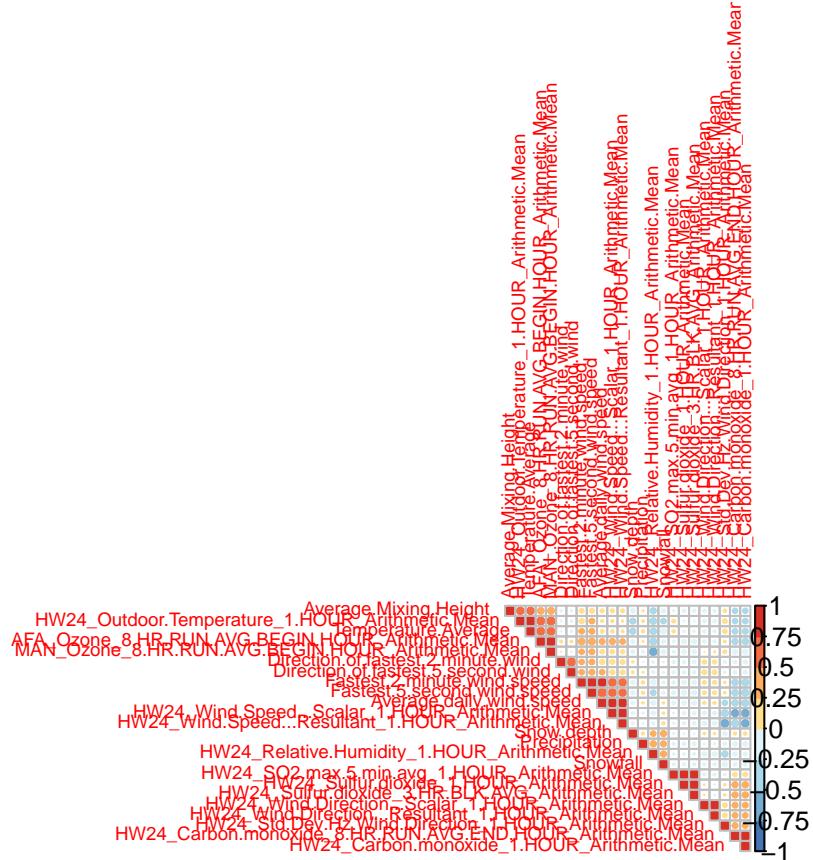
```

#Correlation plot

```

#correlation plot
data_cor = cor(data_filtered[, -c(25)])
corrplot(data_cor, type="upper", order="hclust",
          col=rev(brewer.pal(n=8, name="RdYlBu"))), tl.cex = .6)

```



#Preliminary Models to check VIF

```

#preliminary model for AFA
afa_preliminary_linear_fit = lm(AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR.Arithmetic.Mean ~ .

```

```
-MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean,data=data_filtered)
summary(afa_preliminary_linear_fit) #print summary
```

```
##
## Call:
## lm(formula = AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~
##      . - MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean, data = data_filtered)
##
## Residuals:
##       Min     1Q Median     3Q    Max 
## -0.0270543 -0.0044345 -0.0000031  0.0043759  0.0248062
##
## Coefficients:
##                               Estimate
## (Intercept)                1.786e-02
## HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean      5.459e-03
## HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean   -2.164e-05
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 3.271e-02
## HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean          -3.810e-02
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 1.318e-05
## HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean   -5.153e-03
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean           -2.378e-03
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean    -4.197e-05
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean        7.311e-04
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean         -1.536e-04
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean       6.079e-04
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean -1.592e-04
## Average.Mixing.Height                                4.677e-07
## Average.daily.wind.speed                            -6.438e-05
## Precipitation                                     2.969e-03
## Snowfall                                         1.048e-03
## Snow.depth                                       1.404e-03
## Temperature.Average                           -3.590e-04
## Direction.of.fastest.2.minute.wind            -1.973e-06
## Direction.of.fastest.5.second.wind           -3.096e-06
## Fastest.2.minute.wind.speed                  -1.081e-04
## Fastest.5.second.wind.speed                 1.827e-04
## Date                                            9.921e-07
##
##                               Std. Error t value
## (Intercept)                5.122e-03  3.487
## HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean      1.068e-03  5.112
## HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean   1.521e-05 -1.423
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 4.654e-03  7.028
## HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean          4.892e-03 -7.789
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 1.509e-05  0.874
## HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean   1.143e-03 -4.510
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean           1.143e-03 -2.080
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean    9.959e-04 -0.042
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean        1.776e-04  4.118
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean         1.301e-05 -11.802
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean       5.370e-05  11.321
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean 3.742e-05 -4.255
## Average.Mixing.Height                            1.601e-07  2.922
```

```

## Average.daily.wind.speed          9.115e-05 -0.706
## Precipitation                   1.299e-03  2.286
## Snowfall                        3.184e-04  3.292
## Snow.depth                      2.131e-04  6.592
## Temperature.Average             5.339e-05 -6.725
## Direction.of.fastest.2.minute.wind 1.792e-06 -1.102
## Direction.of.fastest.5.second.wind 1.870e-06 -1.656
## Fastest.2.minute.wind.speed    7.987e-05 -1.354
## Fastest.5.second.wind.speed   5.779e-05  3.161
## Date                            2.419e-07  4.101
##
## Pr(>|t|)
## (Intercept)                     0.000499 ***
## HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean 3.49e-07 ***
## HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean 0.154940
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 2.86e-12 ***
## HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean      1.08e-14 ***
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 0.382396
## HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean 6.87e-06 ***
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean      0.037627 *
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean 0.966388
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean 3.98e-05 ***
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean    < 2e-16 ***
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean  < 2e-16 ***
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean 2.19e-05 ***
## Average.Mixing.Height           0.003516 **
## Average.daily.wind.speed       0.480108
## Precipitation                  0.022339 *
## Snowfall                        0.001012 **
## Snow.depth                      5.56e-11 ***
## Temperature.Average            2.28e-11 ***
## Direction.of.fastest.2.minute.wind 0.270783
## Direction.of.fastest.5.second.wind 0.097910 .
## Fastest.2.minute.wind.speed   0.175984
## Fastest.5.second.wind.speed   0.001594 **
## Date                            4.28e-05 ***
##
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00666 on 1993 degrees of freedom
## Multiple R-squared:  0.5767, Adjusted R-squared:  0.5718
## F-statistic:   118 on 23 and 1993 DF,  p-value: < 2.2e-16

```

```
vif(afa_preliminary_linear_fit) #examine VIF of variables
```

```

##          HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean
##          HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean
##          HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean
##          HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean
##          HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
##          HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean
##          HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean
##          HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean
##          HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean
##          HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
##          HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
##          HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
##          Average.Mixing.Height
##          Average.daily.wind.speed
##          Precipitation
##          Snowfall
##          Snow.depth
##          Temperature.Average
##          Direction.of.fastest.2.minute.wind
##          Direction.of.fastest.5.second.wind
##          Fastest.2.minute.wind.speed
##          Fastest.5.second.wind.speed
##          Date
##          ---
##          Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##          Residual standard error: 0.00666 on 1993 degrees of freedom
##          Multiple R-squared:  0.5767, Adjusted R-squared:  0.5718
##          F-statistic:   118 on 23 and 1993 DF,  p-value: < 2.2e-16

```

```

##          HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean
##                                         160.099824
##          HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean
##                                         90.680665
##          HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean
##                                         78.560303
##          HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean
##                                         15.394565
##          HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
##                                         2.333900
##          HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
##                                         37.036686
##          HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
##                                         5.030765
##          Average.Mixing.Height
##                                         1.532279
##          Average.daily.wind.speed
##                                         4.564613
##          Precipitation
##                                         1.368530
##          Snowfall
##                                         1.399138
##          Snow.depth
##                                         1.290341
##          Temperature.Average
##                                         37.211632
##          Direction.of.fastest.2.minute.wind
##                                         1.810620
##          Direction.of.fastest.5.second.wind
##                                         1.837087
##          Fastest.2.minute.wind.speed
##                                         18.400712
##          Fastest.5.second.wind.speed
##                                         18.459410
##          Date
##                                         1.476583

```

#preliminary model for MAN

```

man_preliminary_linear_fit = lm(MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~ .
                                 -AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean,data=data_filtered)
summary(man_preliminary_linear_fit)

```

```

##
## Call:
## lm(formula = MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~
##     . - AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean, data = data_filtered)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0268463 -0.0041338  0.0001469  0.0040979  0.0244308
##
## Coefficients:
##                               Estimate
## (Intercept)                2.973e-02

```

## HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean	-4.553e-04
## HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean	2.643e-06
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean	-8.102e-03
## HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean	6.242e-03
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean	2.815e-06
## HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean	2.083e-04
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean	-2.208e-03
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetric.Mean	-1.083e-04
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetric.Mean	6.303e-04
## HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean	-2.361e-04
## HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean	5.032e-04
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean	-1.411e-04
## Average.Mixing.Height	4.775e-07
## Average.daily.wind.speed	6.689e-05
## Precipitation	2.142e-03
## Snowfall	6.288e-04
## Snow.depth	1.546e-03
## Temperature.Average	-2.496e-04
## Direction.of.fastest.2.minute.wind	-2.344e-06
## Direction.of.fastest.5.second.wind	-3.052e-06
## Fastest.2.minute.wind.speed	-1.887e-04
## Fastest.5.second.wind.speed	1.941e-04
## Date	8.445e-07
##	
## (Intercept)	Std. Error t value
## HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean	4.912e-03 6.053
## HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean	1.024e-03 -0.445
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean	1.458e-05 0.181
## HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean	4.462e-03 -1.816
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean	4.691e-03 1.331
## HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean	1.447e-05 0.195
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean	1.096e-03 0.190
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetric.Mean	1.096e-03 -2.014
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetric.Mean	9.549e-04 -0.113
## HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean	1.703e-04 3.702
## HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean	1.248e-05 -18.922
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean	5.149e-05 9.774
## Average.Mixing.Height	3.589e-05 -3.931
## Average.daily.wind.speed	1.535e-07 3.111
## Precipitation	8.740e-05 0.765
## Snowfall	1.245e-03 1.720
## Snow.depth	3.053e-04 2.060
## Temperature.Average	2.043e-04 7.569
## Direction.of.fastest.2.minute.wind	5.119e-05 -4.876
## Direction.of.fastest.5.second.wind	1.718e-06 -1.364
## Fastest.2.minute.wind.speed	1.793e-06 -1.702
## Fastest.5.second.wind.speed	7.659e-05 -2.463
## Date	5.542e-05 3.503
##	2.320e-07 3.641
##	Pr(> t)
## (Intercept)	1.70e-09 ***
## HW24_Wind.Speed...Scalar_1.HOUR_Arithmetric.Mean	0.656617
## HW24_Wind.Direction...Scalar_1.HOUR_Arithmetric.Mean	0.856195
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetric.Mean	0.069562 .
## HW24_Carbon.monoxide_1.HOUR_Arithmetric.Mean	0.183478

```

## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean      0.845738
## HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean        0.849245
## HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean                0.044144 *
## HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean         0.909731
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean            0.000220 ***
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean             < 2e-16 ***
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean          < 2e-16 ***
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean    8.76e-05 ***
## Average.Mixing.Height                                    0.001892 **
## Average.daily.wind.speed                                0.444205
## Precipitation                                         0.085577 .
## Snowfall                                              0.039554 *
## Snow.depth                                            5.72e-14 ***
## Temperature.Average                                   1.17e-06 ***
## Direction.of.fastest.2.minute.wind                   0.172623
## Direction.of.fastest.5.second.wind                 0.088849 .
## Fastest.2.minute.wind.speed                         0.013848 *
## Fastest.5.second.wind.speed                        0.000470 ***
## Date                                                 0.000279 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## Residual standard error: 0.006386 on 1993 degrees of freedom
## Multiple R-squared:  0.5674, Adjusted R-squared:  0.5625
## F-statistic: 113.7 on 23 and 1993 DF,  p-value: < 2.2e-16

```

```
vif(man_preliminary_linear_fit)
```

```

##           HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean      149.884855
##           HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean     14.373170
##           HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 17.533377
##           HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean          19.880972
##           HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 13.930858
##           HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean    160.099824
##           HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean           90.680665
##           HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean    78.560303
##           HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean       15.394565
##           HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean        2.333900
##           HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean      37.036686
##           HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean 5.030765
##           Average.Mixing.Height                            1.532279
## 
```

```

## Average.daily.wind.speed
##                               4.564613
## Precipitation
##                               1.368530
## Snowfall
##                               1.399138
## Snow.depth
##                               1.290341
## Temperature.Average
##                               37.211632
## Direction.of.fastest.2.minute.wind
##                               1.810620
## Direction.of.fastest.5.second.wind
##                               1.837087
## Fastest.2.minute.wind.speed
##                               18.400712
## Fastest.5.second.wind.speed
##                               18.459410
## Date
##                               1.476583

```

```
colnames(data_filtered)
```

```

## [1] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [2] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [3] "HW24_Wind.Speed...Scalar_1.HOUR_Arithmetic.Mean"
## [4] "HW24_Wind.Direction...Scalar_1.HOUR_Arithmetic.Mean"
## [5] "HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean"
## [6] "HW24_Carbon.monoxide_1.HOUR_Arithmetic.Mean"
## [7] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean"
## [8] "HW24_Wind.Speed...Resultant_1.HOUR_Arithmetic.Mean"
## [9] "HW24_Sulfur.dioxide_1.HOUR_Arithmetic.Mean"
## [10] "HW24_Sulfur.dioxide_3.HR.BLK.AVG_Arithmetic.Mean"
## [11] "HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean"
## [12] "HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean"
## [13] "HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean"
## [14] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean"
## [15] "Average.Mixing.Height"
## [16] "Average.daily.wind.speed"
## [17] "Precipitation"
## [18] "Snowfall"
## [19] "Snow.depth"
## [20] "Temperature.Average"
## [21] "Direction.of.fastest.2.minute.wind"
## [22] "Direction.of.fastest.5.second.wind"
## [23] "Fastest.2.minute.wind.speed"
## [24] "Fastest.5.second.wind.speed"
## [25] "Date"

```

```
#remove collinear predictors:
```

```

# HW24 Sulfur dioxide 1 hr mean [9] and 3 hour bulk average [10] - retain 5 min avg
# HW24 CO 1_Hour [6],
# COS Avg Temp [20], - retrain HW24 Temp

```

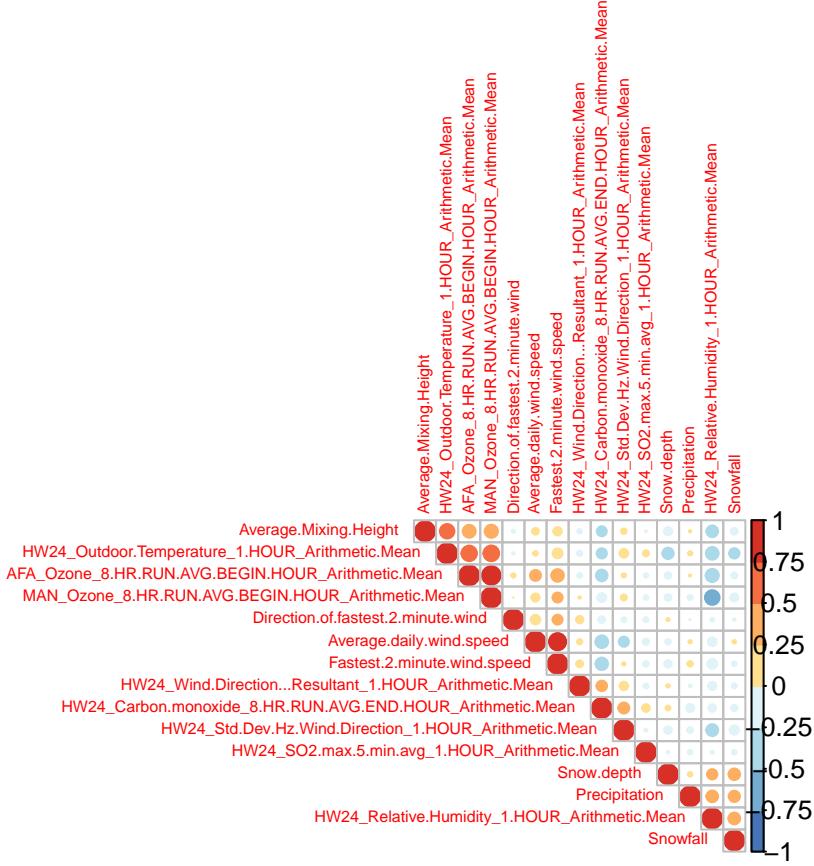
```

# HW24 Scalar wind speed and direction [3,4]
# drop HW24 resultant wind speed [8] - retain COS daily avg wind speed
# COS 5second wind speeds params [22,24]
ozone_8_hr = data_filtered[,-c(3,4,6,8,9,10,20,22,24)]
colnames(ozone_8_hr)

## [1] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetric.Mean"
## [2] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetric.Mean"
## [3] "HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetric.Mean"
## [4] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetric.Mean"
## [5] "HW24_S02.max.5.min.avg_1.HOUR_Arithmetric.Mean"
## [6] "HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean"
## [7] "HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean"
## [8] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean"
## [9] "Average.Mixing.Height"
## [10] "Average.daily.wind.speed"
## [11] "Precipitation"
## [12] "Snowfall"
## [13] "Snow.depth"
## [14] "Direction.of.fastest.2.minute.wind"
## [15] "Fastest.2.minute.wind.speed"
## [16] "Date"

#new correlation plot
data_cor = cor(ozone_8_hr[,-c(16)]) #don't include date for corrplot
corrplot(data_cor, type="upper", order="hclust",
          col=rev(brewer.pal(n=8, name="RdYlBu"))), tl.cex = .5)

```



#Preliminary Model to check VIF between Fasted 2 minute wind speed and Average Daily Wind Speed

```
afa_preliminary_linear_fit = lm(AFA_Ozone_8.HR.RUN.AVG.BEGIN.HOUR_Arithmetric.Mean ~ .
                                 -MAN_Ozone_8.HR.RUN.AVG.BEGIN.HOUR_Arithmetric.Mean,data=ozone_8_hr)
summary(afa_preliminary_linear_fit)
```

```
##
## Call:
## lm(formula = AFA_Ozone_8.HR.RUN.AVG.BEGIN.HOUR_Arithmetric.Mean ~
##     . - MAN_Ozone_8.HR.RUN.AVG.BEGIN.HOUR_Arithmetric.Mean, data = ozone_8_hr)
##
## Residuals:
##      Min        1Q        Median        3Q       Max 
## -0.0222616 -0.0049559 -0.0000257  0.0049000  0.0246899 
## 
## Coefficients:
##                               Estimate
## (Intercept)                2.500e-02
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetric.Mean -1.043e-02
## HW24_Wind.Direction...Resultant.1.HOUR_Arithmetric.Mean      -1.329e-05
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetric.Mean             -1.292e-04
## HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean            -1.806e-04
## HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean           2.745e-04
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean    -7.995e-05
## Average.Mixing.Height                                3.351e-07
## Average.daily.wind.speed                            9.847e-05
```

```

## Precipitation           2.186e-03
## Snowfall                1.129e-03
## Snow.depth              1.653e-03
## Direction.of.fastest.2.minute.wind -1.990e-06
## Fastest.2.minute.wind.speed      1.982e-04
## Date                      5.789e-07
##
##                               Std. Error t value
## (Intercept)                  4.974e-03  5.026
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 1.618e-03 -6.450
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean     4.715e-06 -2.819
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean            5.427e-05 -2.381
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean            1.305e-05 -13.838
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean          1.313e-05  20.905
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean    2.210e-05 -3.618
## Average.Mixing.Height          1.671e-07  2.006
## Average.daily.wind.speed       7.994e-05  1.232
## Precipitation                 1.350e-03  1.620
## Snowfall                      3.321e-04  3.399
## Snow.depth                     2.194e-04  7.534
## Direction.of.fastest.2.minute.wind 1.471e-06 -1.352
## Fastest.2.minute.wind.speed     3.393e-05  5.842
## Date                          2.349e-07  2.464
##
##                               Pr(>|t|)
## (Intercept)                  5.44e-07 ***
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 1.40e-10 ***
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean     0.004862 **
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean            0.017361 *
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean            < 2e-16 ***
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean          < 2e-16 ***
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean    0.000304 ***
## Average.Mixing.Height          0.044982 *
## Average.daily.wind.speed       0.218135
## Precipitation                 0.105399
## Snowfall                      0.000689 ***
## Snow.depth                     7.41e-14 ***
## Direction.of.fastest.2.minute.wind 0.176422
## Fastest.2.minute.wind.speed     6.01e-09 ***
## Date                          0.013822 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ',' 1
##
## Residual standard error: 0.007009 on 2002 degrees of freedom
## Multiple R-squared:  0.529, Adjusted R-squared:  0.5257
## F-statistic: 160.6 on 14 and 2002 DF, p-value: < 2.2e-16

vif(afa_preliminary_linear_fit)

```

```

## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean
##                                         1.913186
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
##                                         1.228544
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean
##                                         1.298516
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean

```

```

##                                     2.120567
##      HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
##                                         1.999220
##      HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
##                                         1.583818
##      Average.Mixing.Height
##                                         1.507143
##      Average.daily.wind.speed
##                                         3.169850
##      Precipitation
##                                         1.334794
##      Snowfall
##                                         1.374248
##      Snow.depth
##                                         1.235529
##      Direction.of.fastest.2.minute.wind
##                                         1.102674
##      Fastest.2.minute.wind.speed
##                                         2.997651
##      Date
##                                         1.257255

man_preliminary_linear_fit = lm(MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~ .
                                 -AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean,data=ozone_8_hr)
summary(man_preliminary_linear_fit)

##
## Call:
## lm(formula = MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~
##     . - AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean, data = ozone_8_hr)
##
## Residuals:
##      Min        1Q        Median         3Q        Max
## -0.022752 -0.004162  0.000046  0.004158  0.025145
##
## Coefficients:
##                               Estimate
## (Intercept)                3.058e-02
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean -6.172e-03
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean      5.734e-06
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean            -1.987e-04
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean            -2.426e-04
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean           2.720e-04
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean      -1.345e-04
## Average.Mixing.Height                                4.199e-07
## Average.daily.wind.speed                            -8.988e-05
## Precipitation                                1.684e-03
## Snowfall                                    5.612e-04
## Snow.depth                                  1.742e-03
## Direction.of.fastest.2.minute.wind          -3.099e-06
## Fastest.2.minute.wind.speed                  4.180e-05
## Date                                       8.043e-07
## Std. Error t value
## (Intercept) 4.600e-03   6.648

```

```

## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 1.496e-03 -4.124
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean      4.361e-06  1.315
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean            5.020e-05 -3.958
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean             1.208e-05 -20.088
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean          1.214e-05 22.395
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean    2.044e-05 -6.581
## Average.Mixing.Height                                    1.545e-07  2.717
## Average.daily.wind.speed                                7.394e-05 -1.216
## Precipitation                                         1.248e-03  1.349
## Snowfall                                              3.072e-04  1.827
## Snow.depth                                            2.029e-04  8.584
## Direction.of.fastest.2.minute.wind                   1.361e-06 -2.277
## Fastest.2.minute.wind.speed                           3.138e-05  1.332
## Date                                                 2.173e-07  3.701
##
## Pr(>|t|)
## (Intercept)                                         3.82e-11 ***
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 3.87e-05 ***
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean      0.188685
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean            7.82e-05 ***
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean             < 2e-16 ***
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean          < 2e-16 ***
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean    5.95e-11 ***
## Average.Mixing.Height                                    0.006645 **
## Average.daily.wind.speed                                0.224303
## Precipitation                                         0.177448
## Snowfall                                              0.067866 .
## Snow.depth                                            < 2e-16 ***
## Direction.of.fastest.2.minute.wind                   0.022897 *
## Fastest.2.minute.wind.speed                           0.183076
## Date                                                 0.000221 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.006483 on 2002 degrees of freedom
## Multiple R-squared:  0.5522, Adjusted R-squared:  0.5491
## F-statistic: 176.3 on 14 and 2002 DF,  p-value: < 2.2e-16

```

```
vif(man_preliminary_linear_fit)
```

```

## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 1.913186
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean      1.228544
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean            1.298516
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean             2.120567
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean          1.999220
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean    1.583818
## Average.Mixing.Height                                    1.507143
## Average.daily.wind.speed

```

```

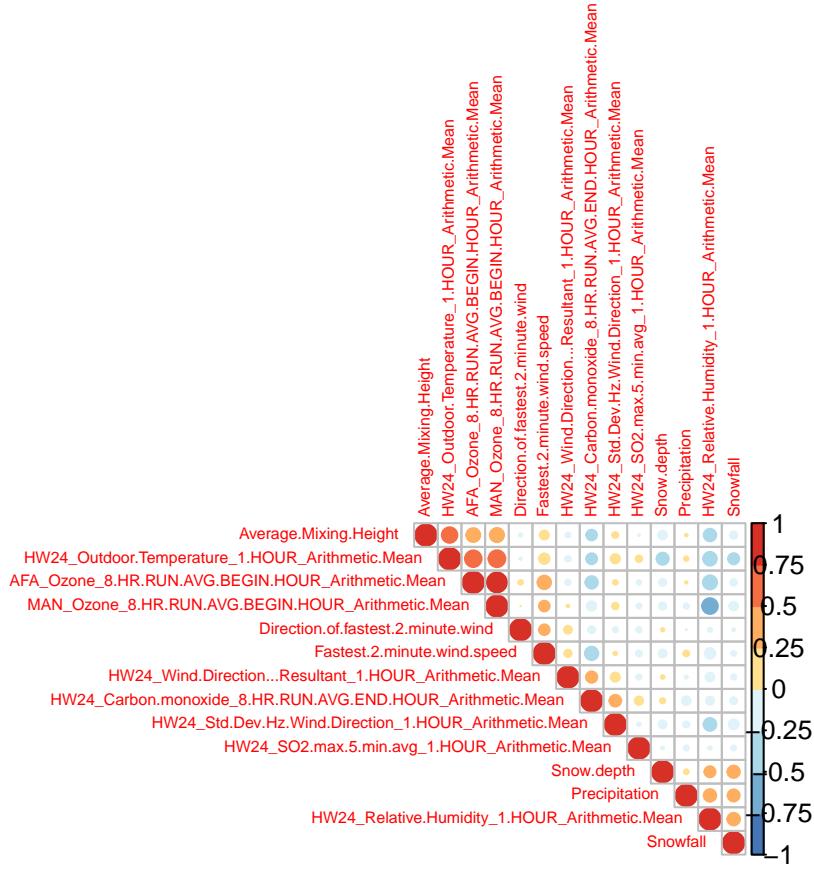
##          3.169850
##          Precipitation
##          1.334794
##          Snowfall
##          1.374248
##          Snow.depth
##          1.235529
##          Direction.of.fastest.2.minute.wind
##          1.102674
##          Fastest.2.minute.wind.speed
##          2.997651
##          Date
##          1.257255

#remove Average Daily Wind Speed due to higher vif and collinearity with fastest 2 minute wind speed
ozone_8_hr = ozone_8_hr[,-c(10)]
colnames(ozone_8_hr)

## [1] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [2] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [3] "HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean"
## [4] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean"
## [5] "HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean"
## [6] "HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean"
## [7] "HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean"
## [8] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean"
## [9] "Average.Mixing.Height"
## [10] "Precipitation"
## [11] "Snowfall"
## [12] "Snow.depth"
## [13] "Direction.of.fastest.2.minute.wind"
## [14] "Fastest.2.minute.wind.speed"
## [15] "Date"

#new correlation plot
data_cor = cor(ozone_8_hr[,-c(15)]) #don't include date for corrplot
corrplot(data_cor, type="upper", order="hclust",
         col=rev(brewer.pal(n=8, name="RdYlBu"))), tl.cex = .5)

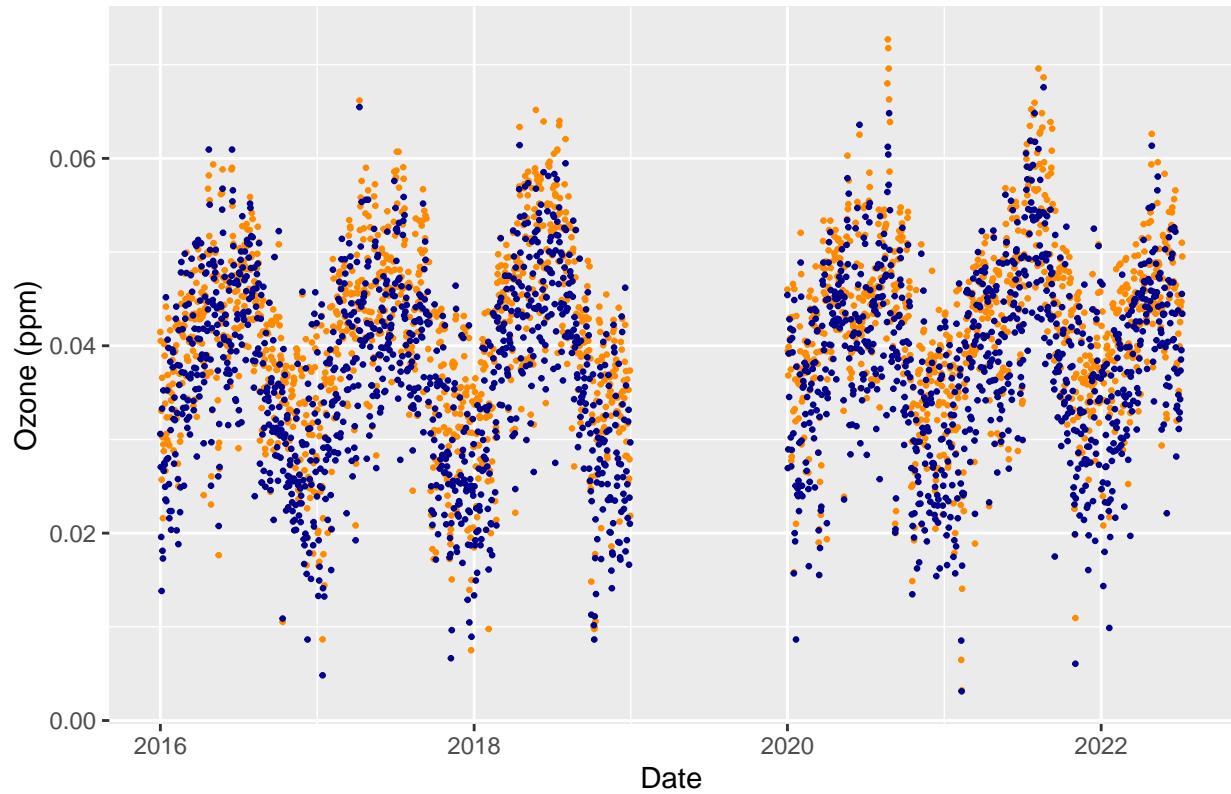
```



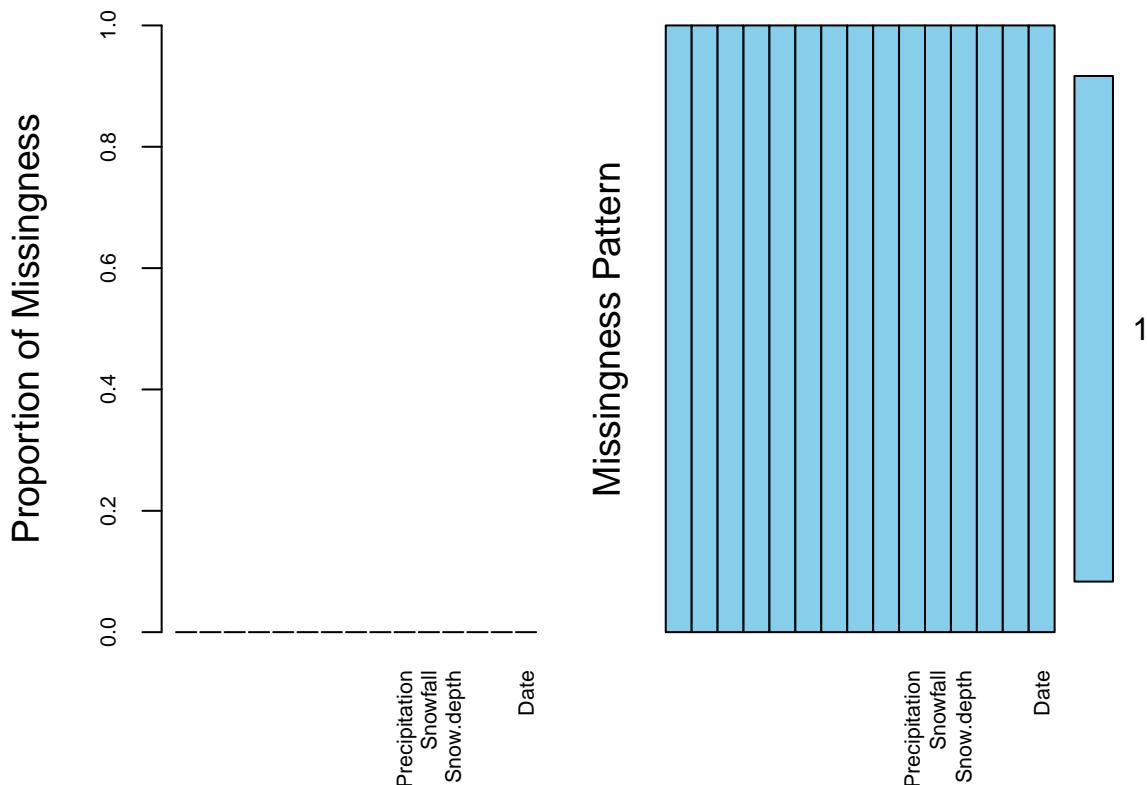
#Explore data and plot MAN_8_hr over time

```
#MAN: orange, AFA: blue (Air Force colors)
ggplot(ozone_8_hr, aes(y=MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean, x=Date), ylab("Ozone")) +
```

Ozone Concentrations at AFA and MAN Monitors by Date



```
#Ensure no missingness  
  
aggr(ozone_8_hr, numbers=TRUE, sortVars=TRUE, cex.axis=.7, gap=3,  
      ylab=c("Proportion of Missingness", "Missingness Pattern"))
```

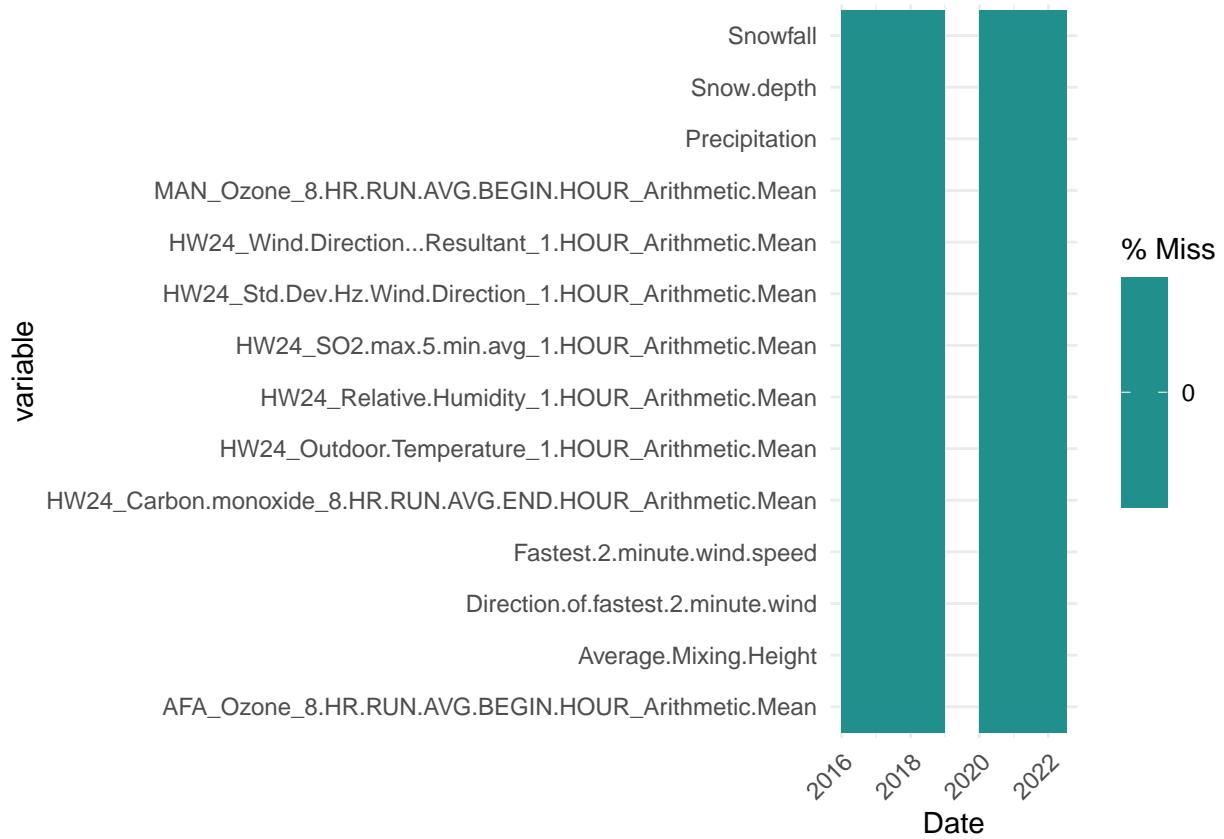


```

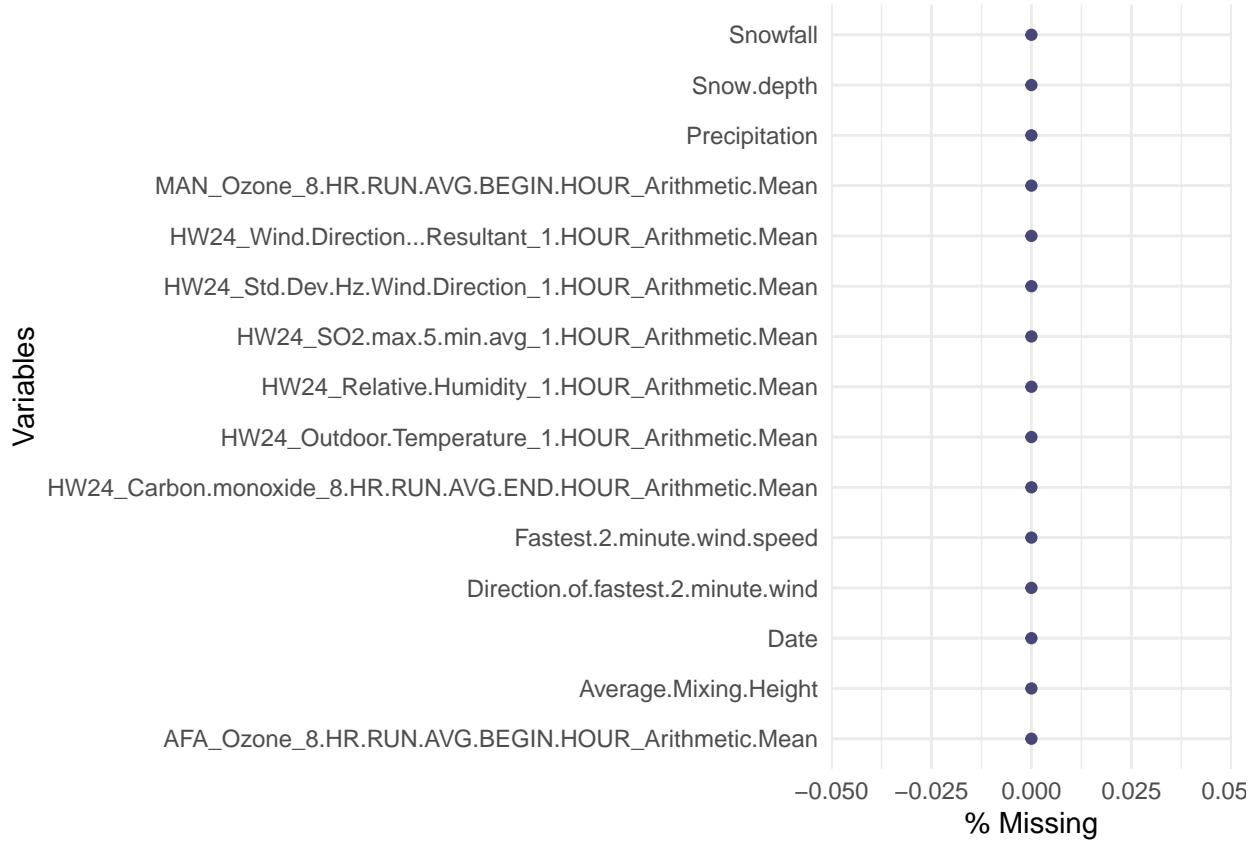
## 
##  Variables sorted by number of missings:
##                                         Variable Count
##          AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetric.Mean      0
##          MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetric.Mean      0
##          HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetric.Mean      0
##          HW24_Wind.Direction...Resultant_1.HOUR_Arithmetric.Mean      0
##          HW24_SO2.max.5.min.avg_1.HOUR_Arithmetric.Mean      0
##          HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean      0
##          HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean      0
##          HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean      0
##          Average.Mixing.Height      0
##          Precipitation      0
##          Snowfall      0
##          Snow.depth      0
##          Direction.of.fastest.2.minute.wind      0
##          Fastest.2.minute.wind.speed      0
##          Date      0

gg_miss_fct(ozone_8_hr, Date)

```



```
gg_miss_var(ozone_8_hr, show_pct = TRUE)
```



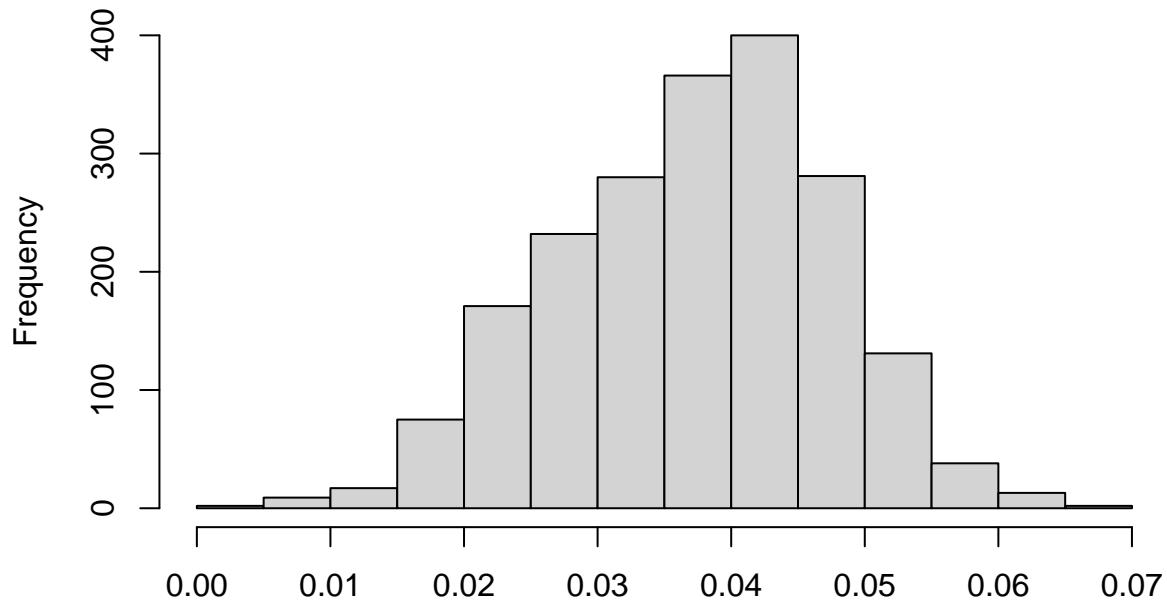
```
colnames(ozone_8_hr)
```

```
## [1] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [2] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
## [3] "HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean"
## [4] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean"
## [5] "HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean"
## [6] "HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean"
## [7] "HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean"
## [8] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean"
## [9] "Average.Mixing.Height"
## [10] "Precipitation"
## [11] "Snowfall"
## [12] "Snow.depth"
## [13] "Direction.of.fastest.2.minute.wind"
## [14] "Fastest.2.minute.wind.speed"
## [15] "Date"
```

Check linear modeling assumptions

```
#examine histograms
hist(as.numeric(ozone_8_hr$AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean)) # relatively normal
```

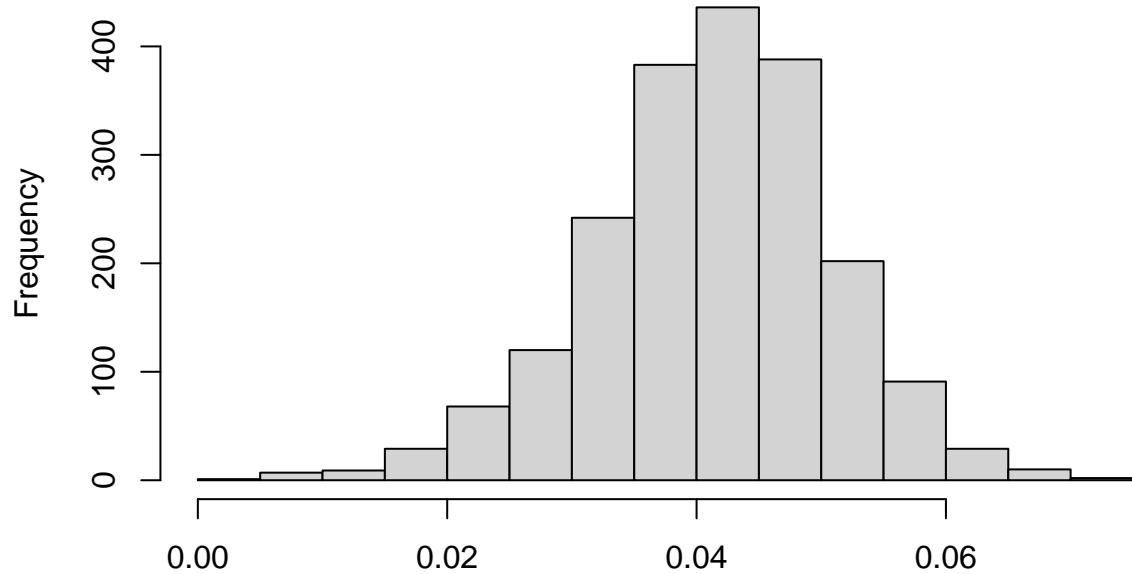
```
if as.numeric(ozone_8_hr$AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_A
```



```
as.numeric(ozone_8_hr$AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mea
```

```
hist(as.numeric(ozone_8_hr$MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean)) # relatively normal
```

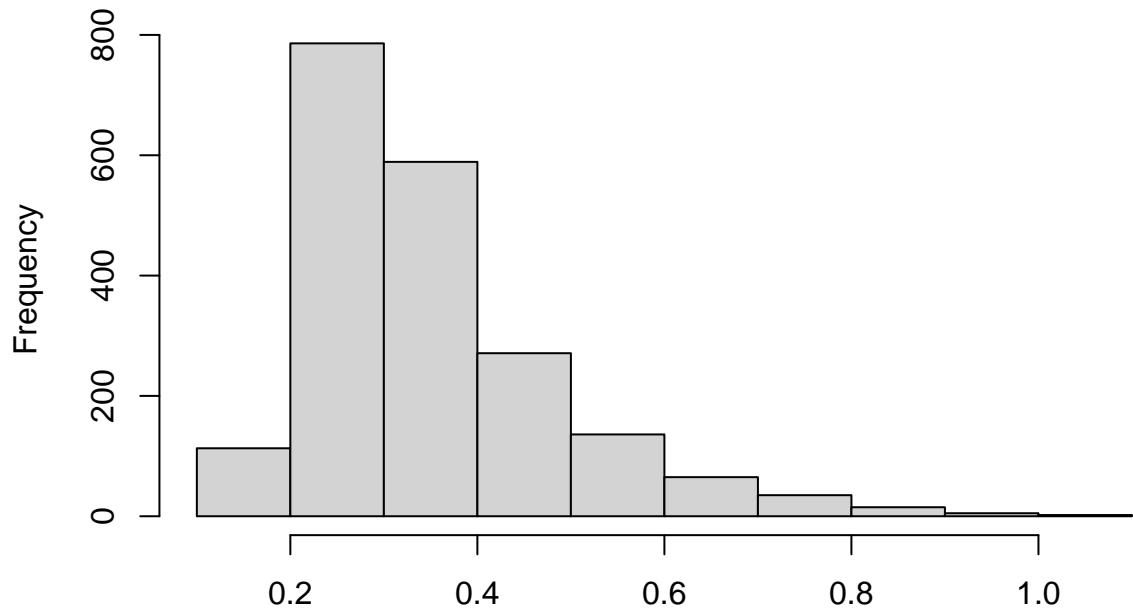
```
f as.numeric(ozone_8_hr$MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_A
```



```
as.numeric(ozone_8_hr$MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mea
```

```
hist(as.numeric(ozone_8_hr$HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean)) # ----->non-
```

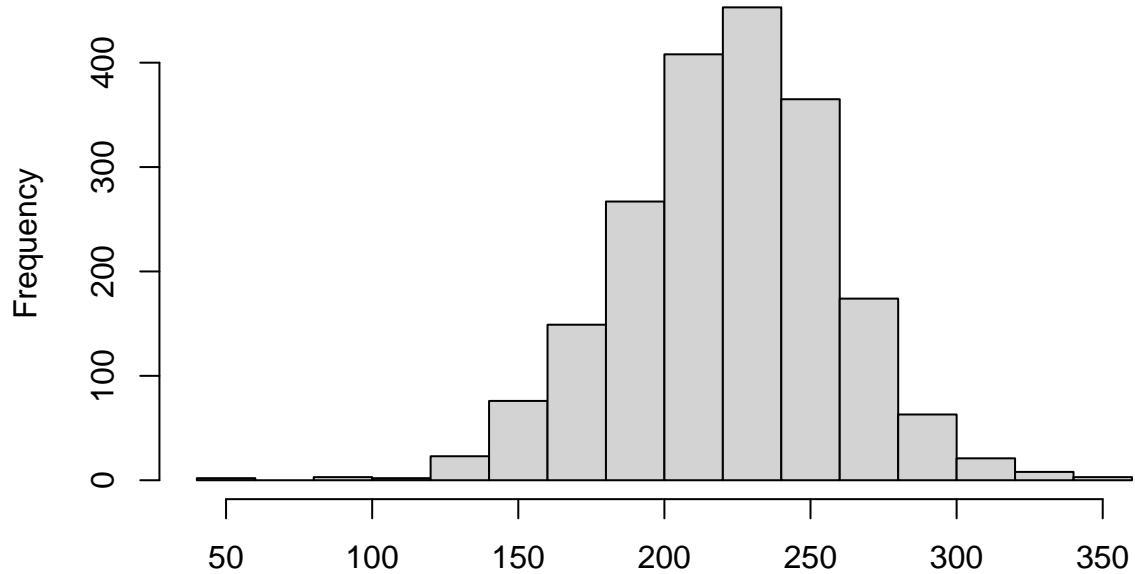
```
as.numeric(ozone_8_hr$HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithme
```



```
.numeric(ozone_8_hr$HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithme
```

```
hist(as.numeric(ozone_8_hr$HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean)) # relatively normal
```

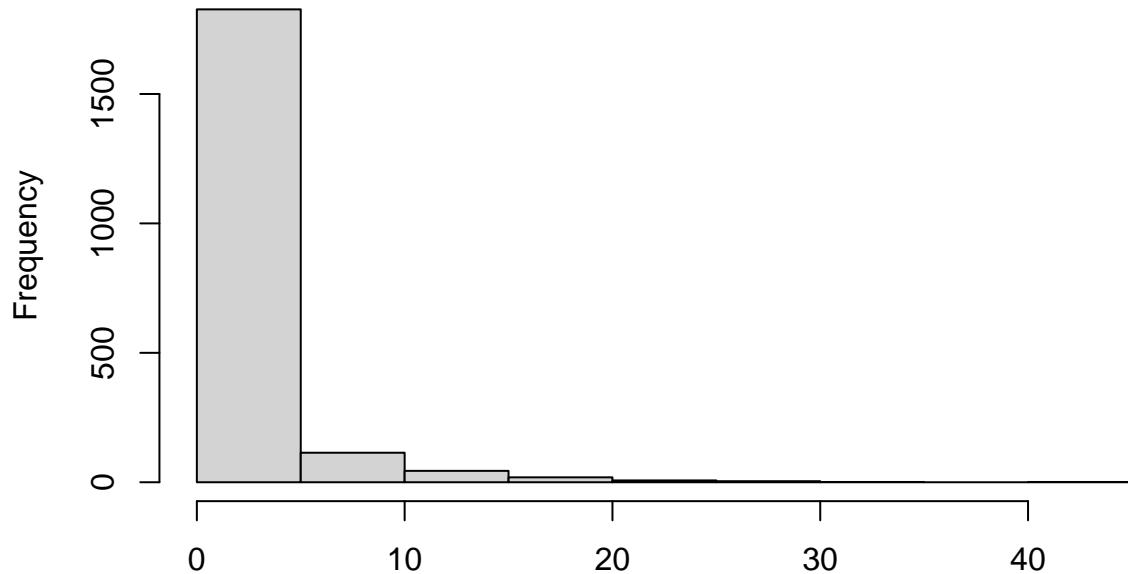
```
f as.numeric(ozone_8_hr$HW24_Wind.Direction...Resultant_1.HOUR_A
```



```
as.numeric(ozone_8_hr$HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mea
```

```
hist(as.numeric(ozone_8_hr$HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean)) # ----->non-normal
```

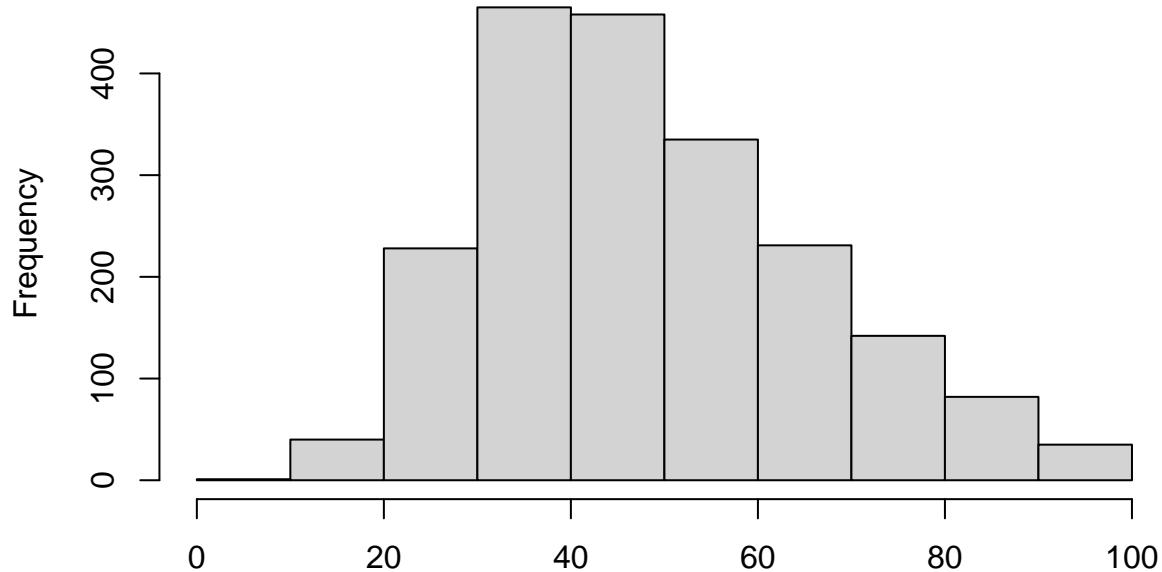
m of as.numeric(ozone_8_hr\$HW24_SO2.max.5.min.avg_1.HOUR_Aritl



as.numeric(ozone_8_hr\$HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean)

```
hist(as.numeric(ozone_8_hr$HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean)) # relatively normal
```

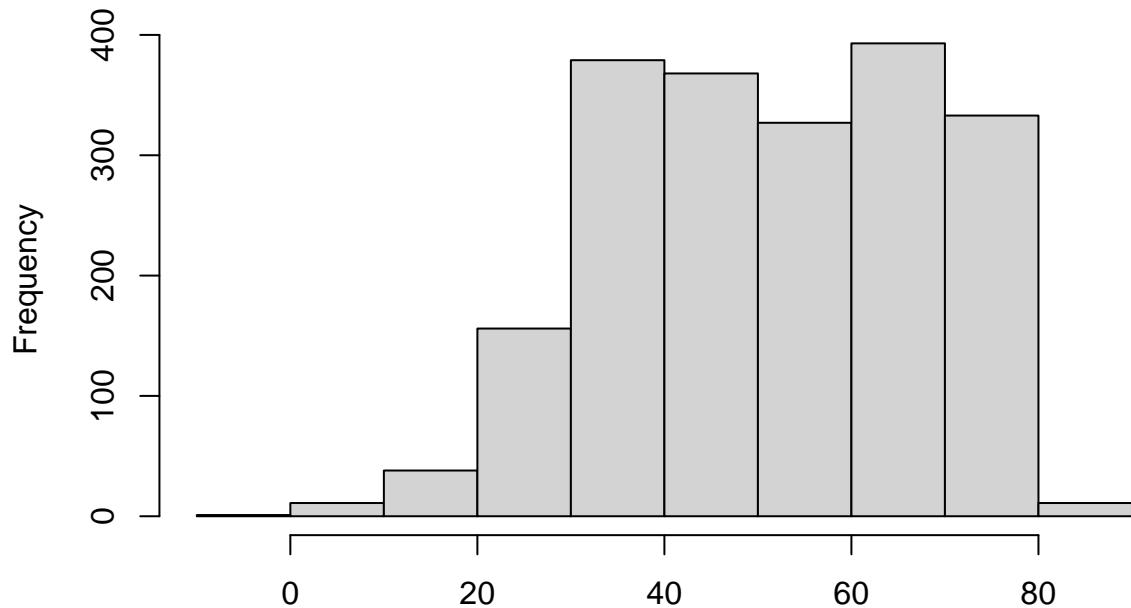
```
#am of as.numeric(ozone_8_hr$HW24_Relative.Humidity_1.HOUR_Arith
```



```
as.numeric(ozone_8_hr$HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean)
```

```
hist(as.numeric(ozone_8_hr$HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean)) # relatively normal
```

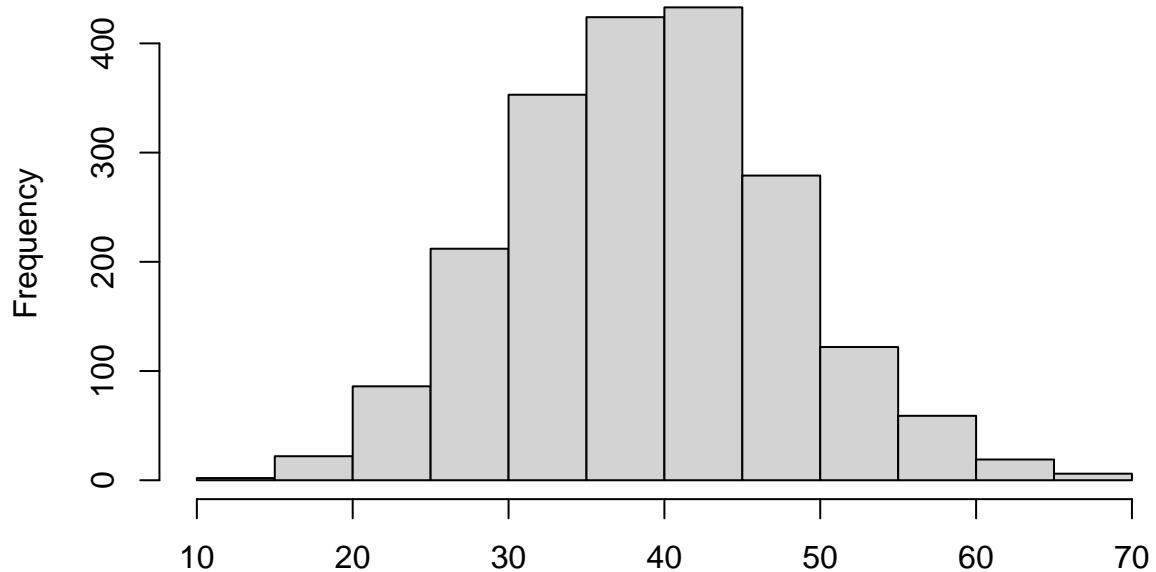
```
1 of as.numeric(ozone_8_hr$HW24_Outdoor.Temperature_1.HOUR_Arithme
```



```
as.numeric(ozone_8_hr$HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean)
```

```
hist(as.numeric(ozone_8_hr$HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean)) # relatively normal
```

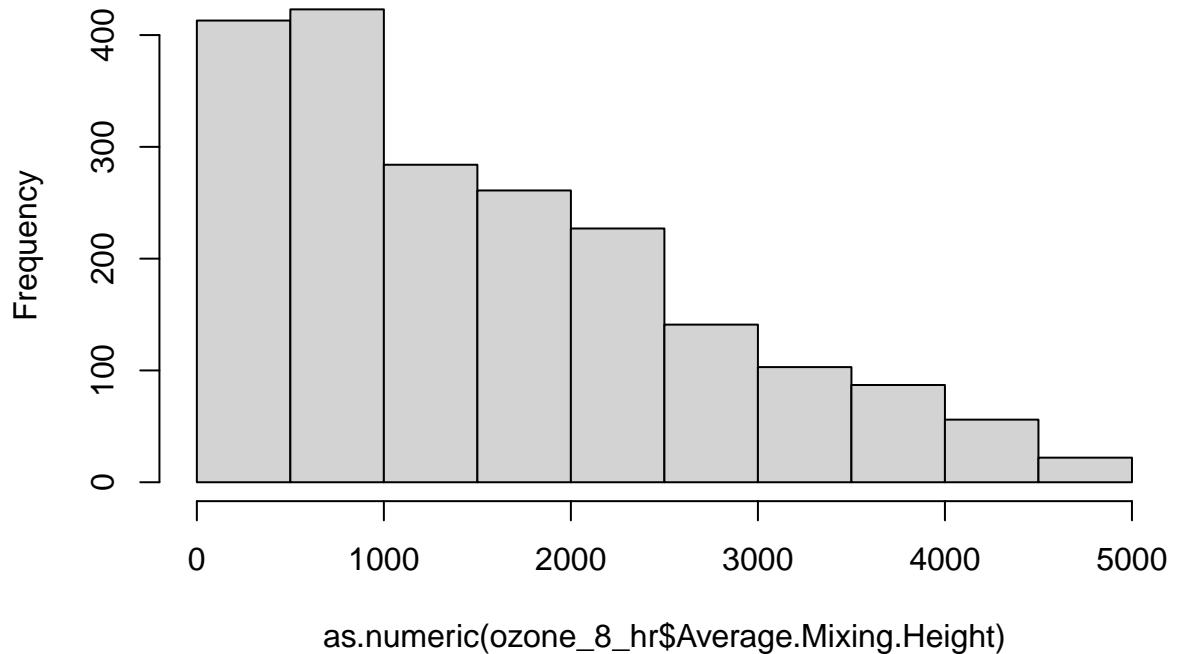
```
f as.numeric(ozone_8_hr$HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_A
```



```
as.numeric(ozone_8_hr$HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mea
```

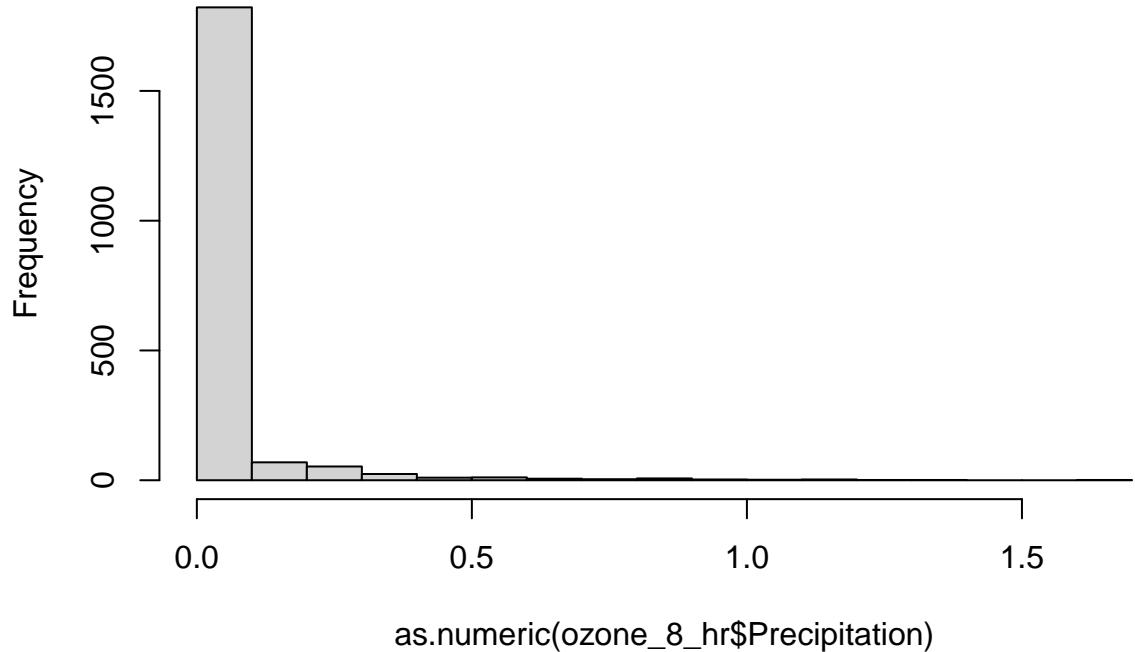
```
hist(as.numeric(ozone_8_hr$Average.Mixing.Height)) # ----->non-normal
```

Histogram of as.numeric(ozone_8_hr\$Average.Mixing.Height)

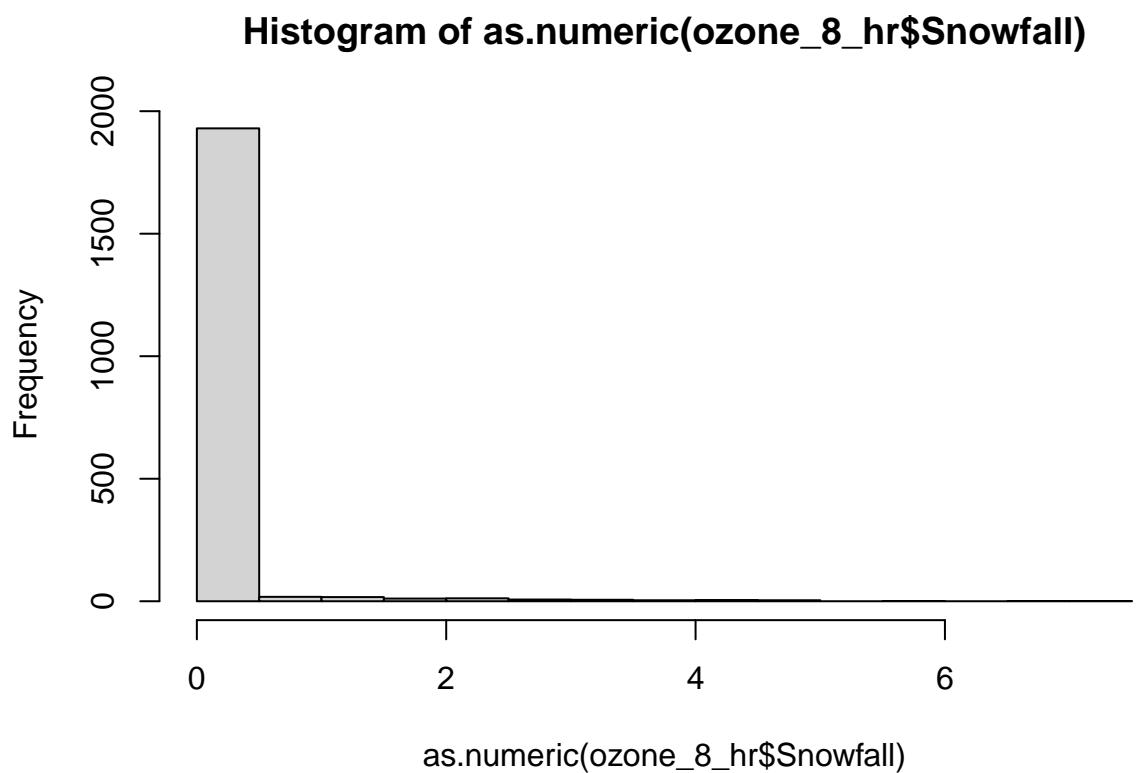


```
hist(as.numeric(ozone_8_hr$Precipitation)) # ----->unbalanced
```

Histogram of as.numeric(ozone_8_hr\$Precipitation)

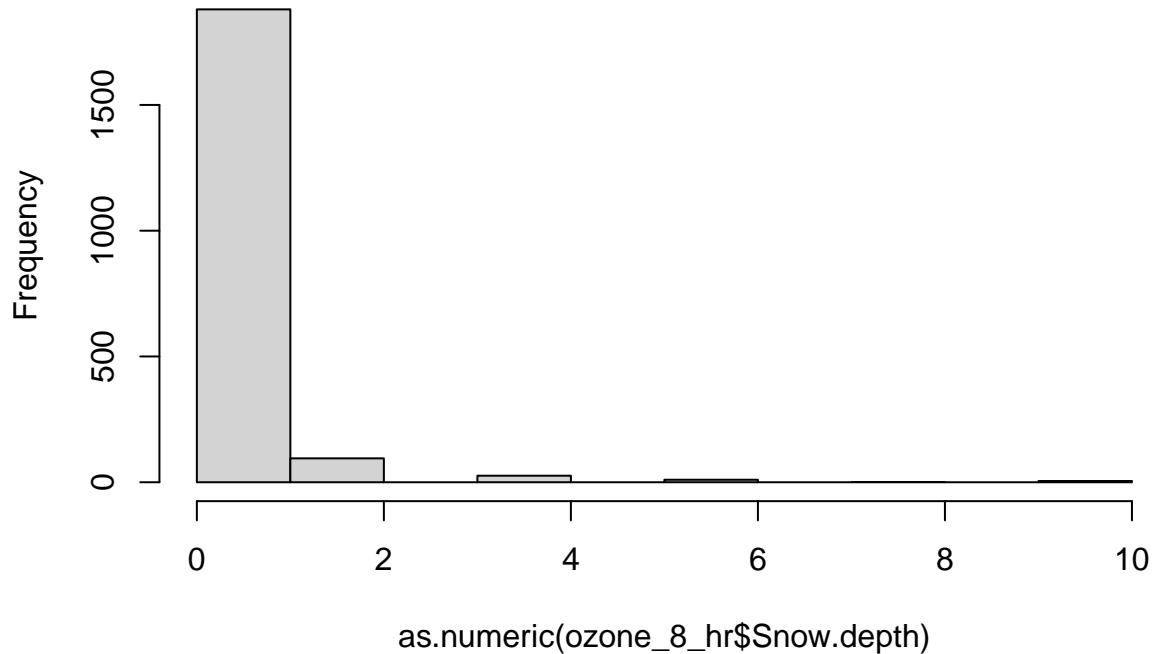


```
hist(as.numeric(ozone_8_hr$Snowfall)) # ----->unbalanced
```



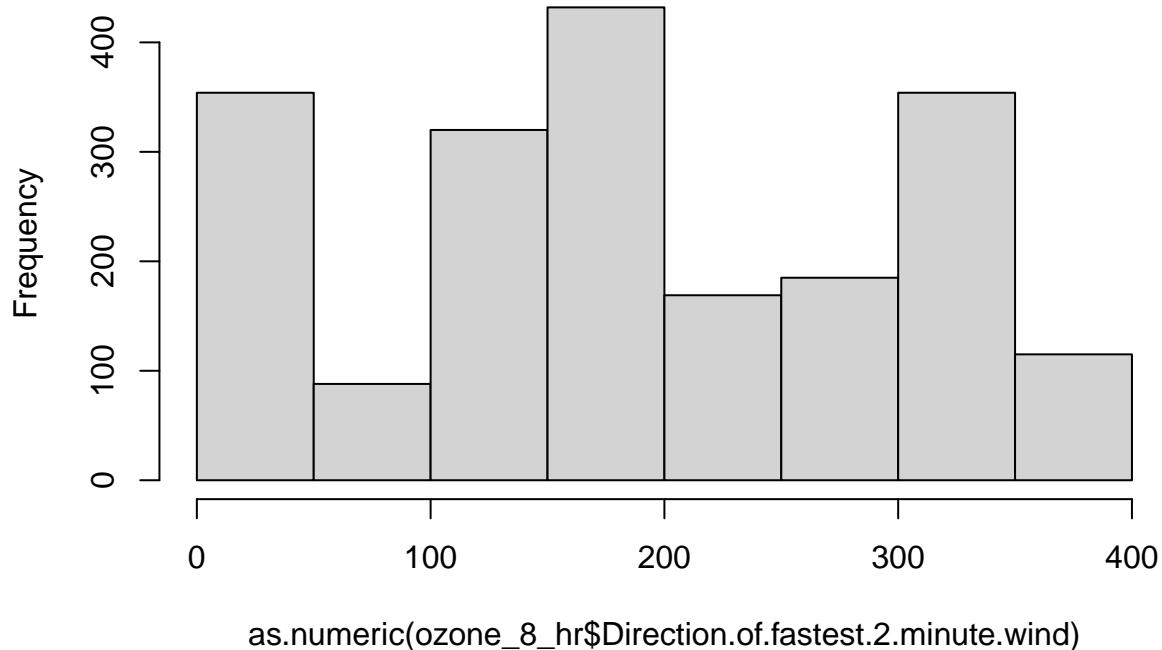
```
hist(as.numeric(ozone_8_hr$Snow.depth)) # ----->unbalanced
```

Histogram of as.numeric(ozone_8_hr\$Snow.depth)



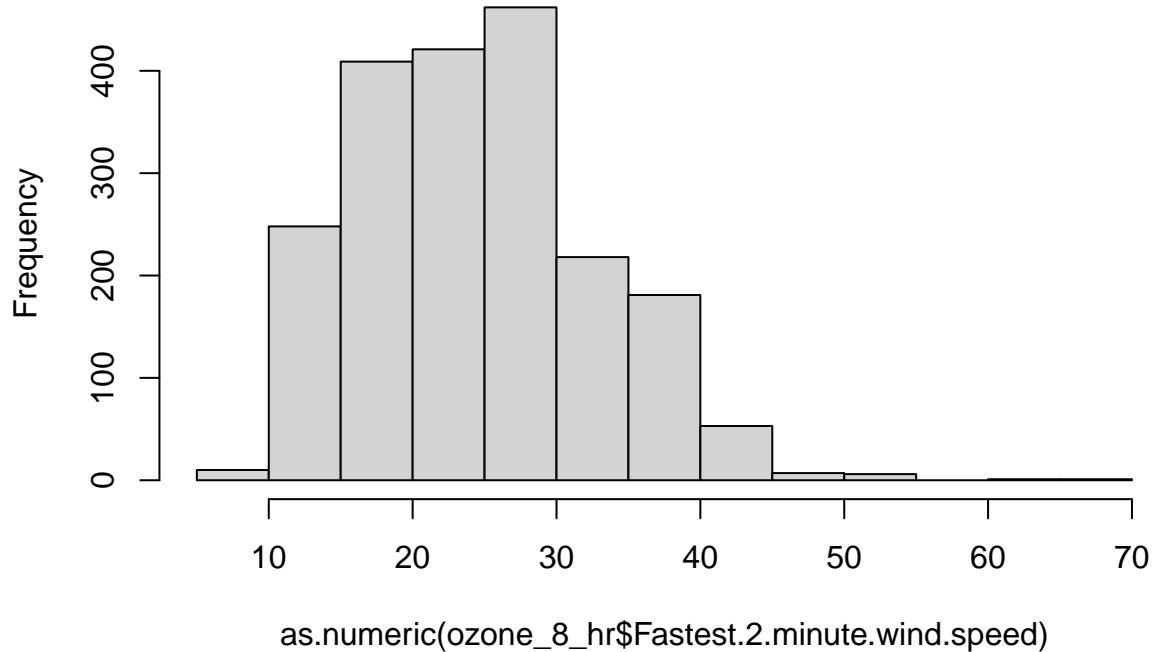
```
hist(as.numeric(ozone_8_hr$Direction.of.fastest.2.minute.wind)) # relatively normal
```

Histogram of as.numeric(ozone_8_hr\$Direction.of.fastest.2.minute.wi

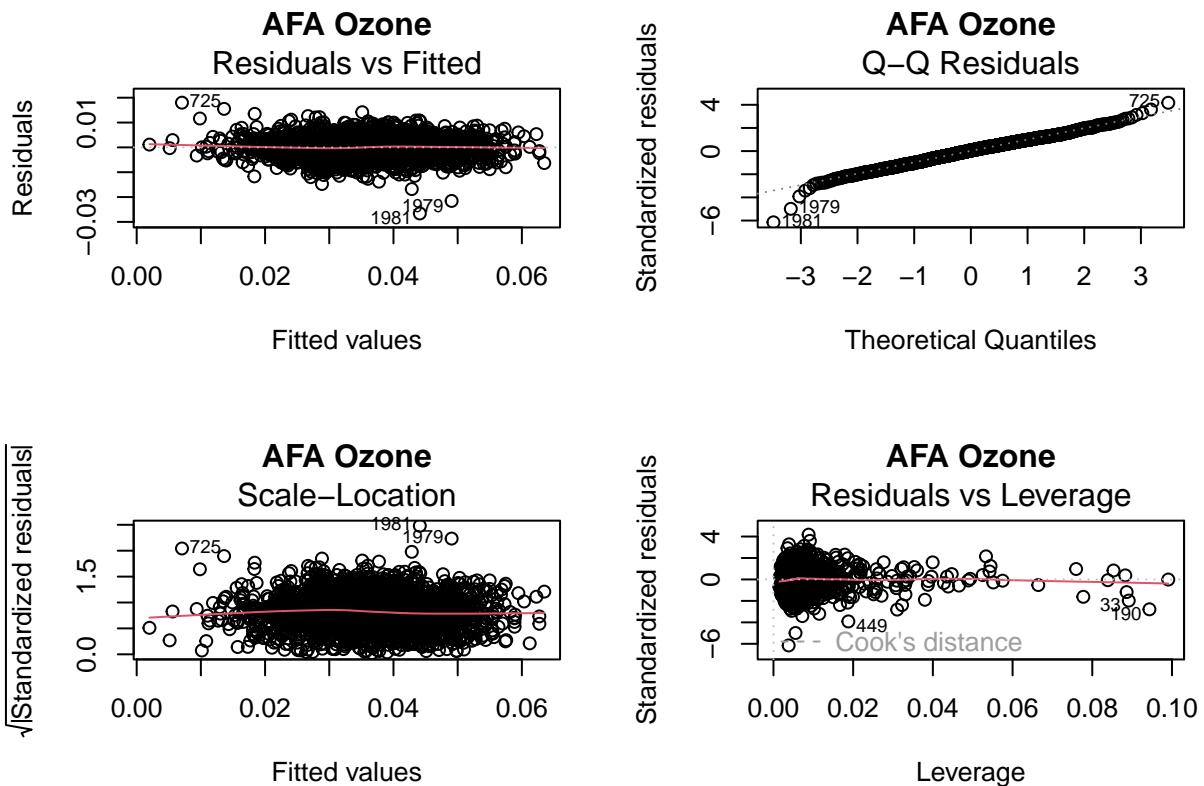


```
hist(as.numeric(ozone_8_hr$Fastest.2.minute.wind.speed)) # relatively normal
```

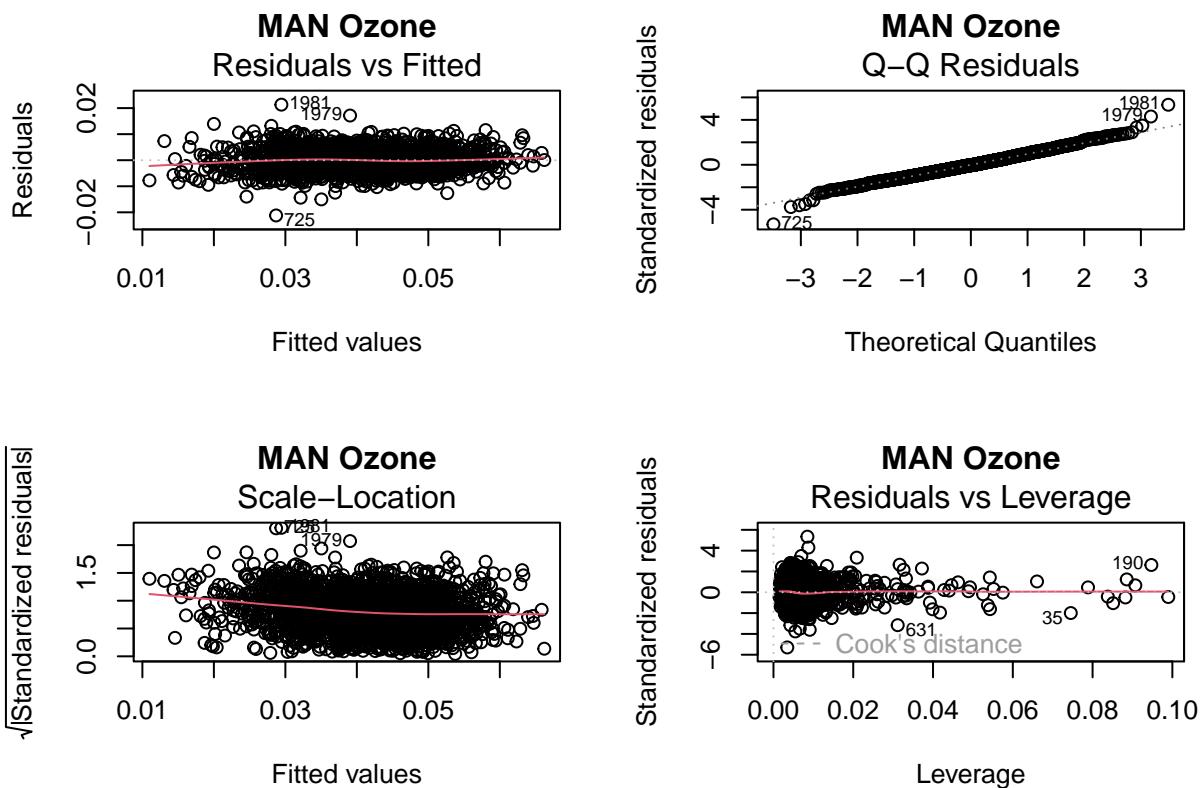
Histogram of as.numeric(ozone_8_hr\$Fastest.2.minute.wind.speed)



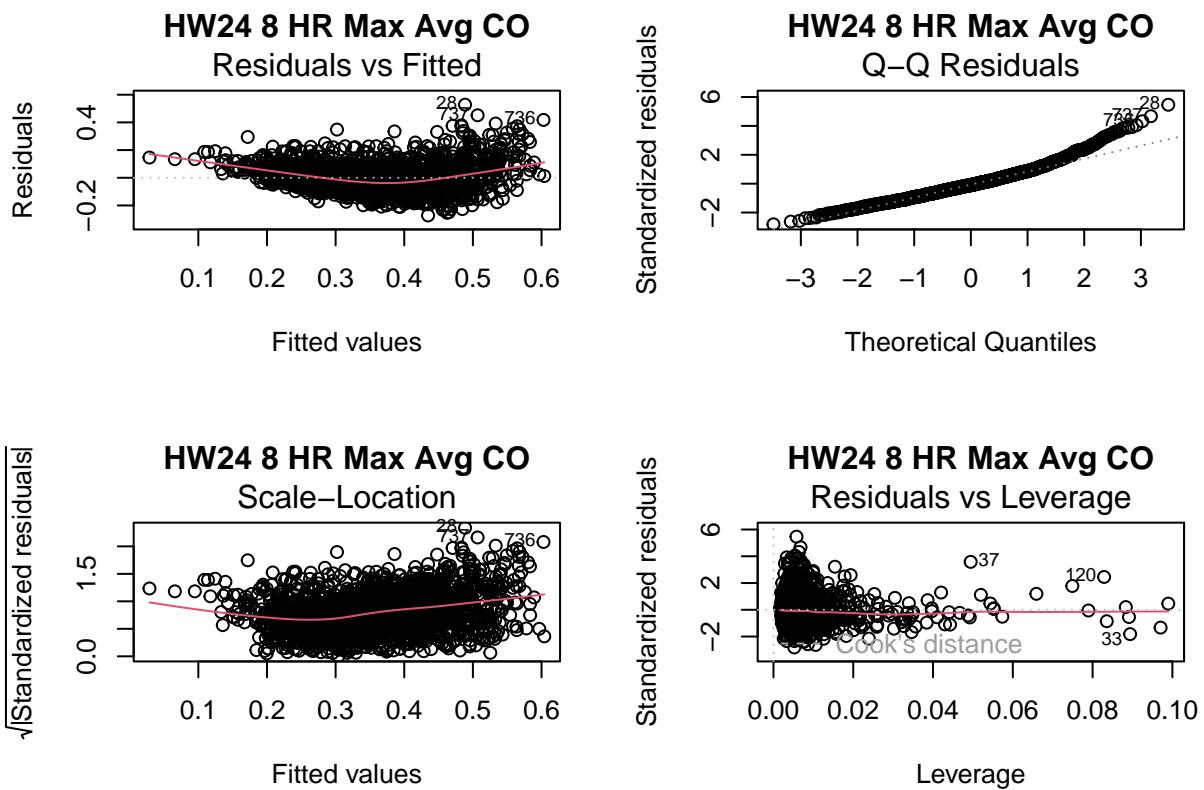
```
#AFA ozone
mod = lm(AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="AFA Ozone")
```



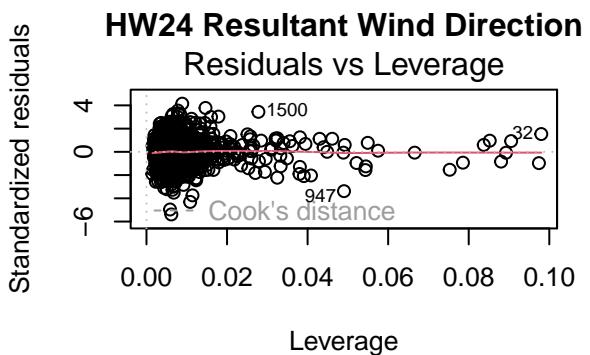
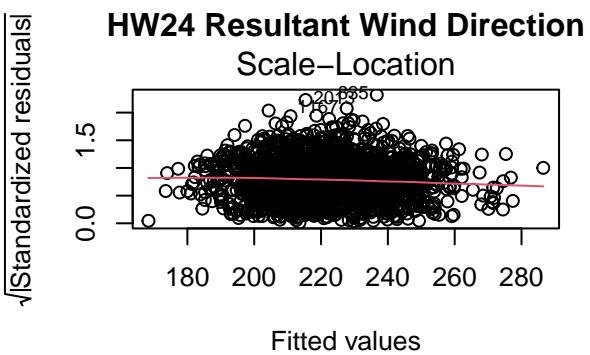
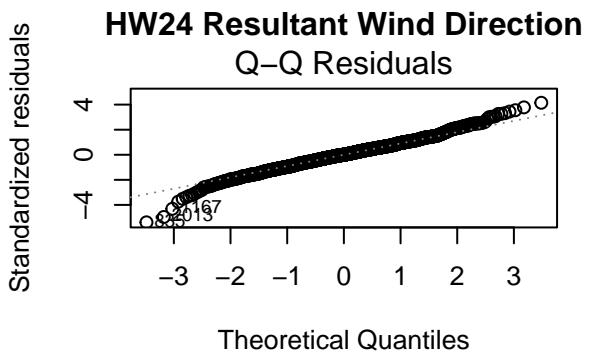
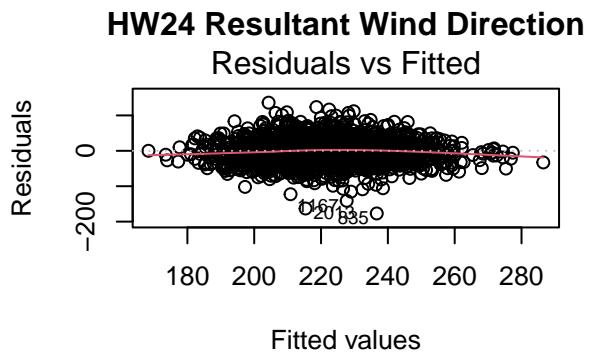
```
#MAN ozone
mod = lm(MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="MAN Ozone")
```



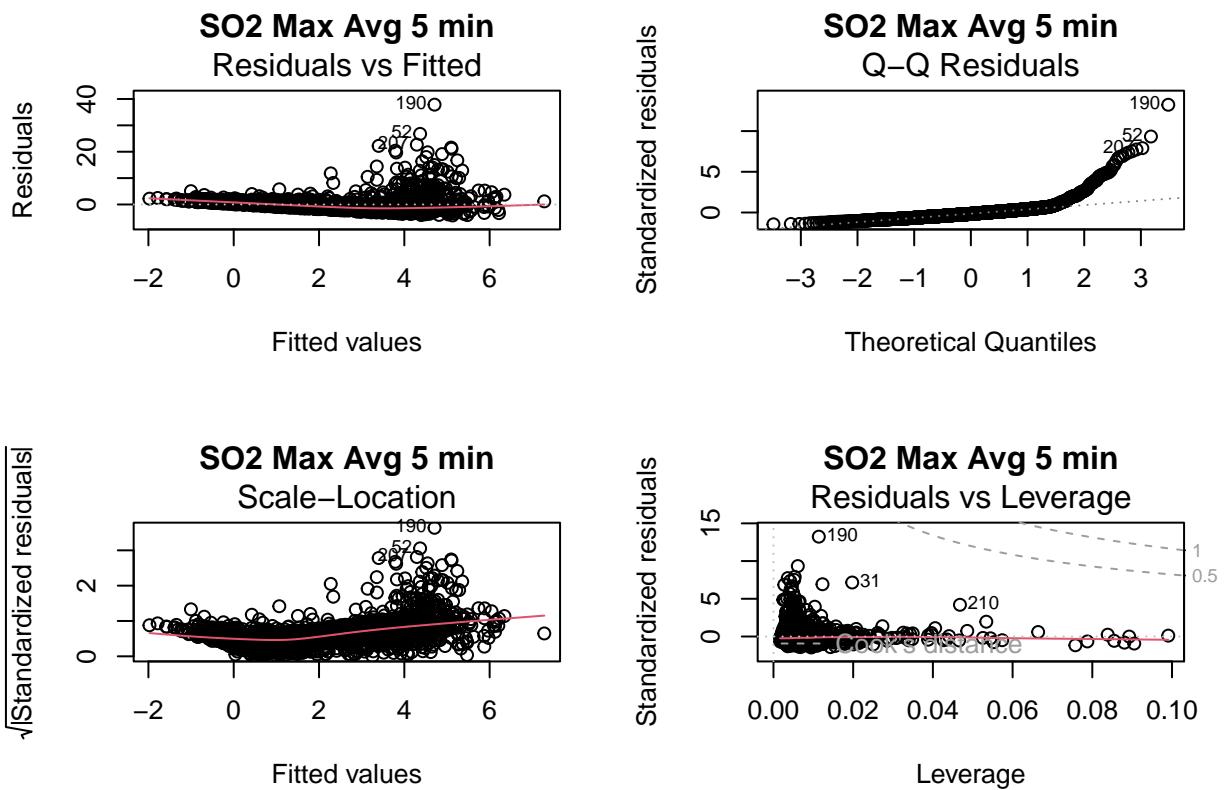
```
#HW24 CO
mod = lm(HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="HW24 8 HR Max Avg CO")
```



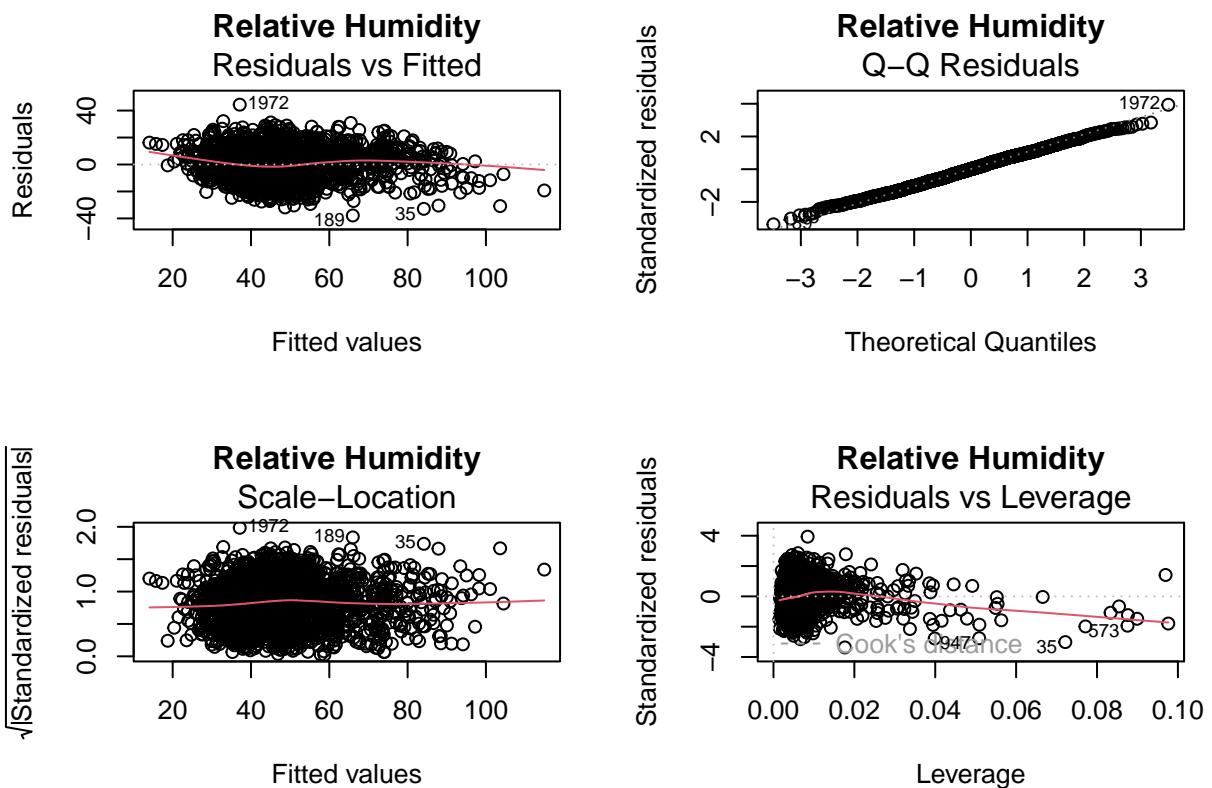
```
#HW24 Wind direction
mod = lm(HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean~, 
         data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="HW24 Resultant Wind Direction")
```



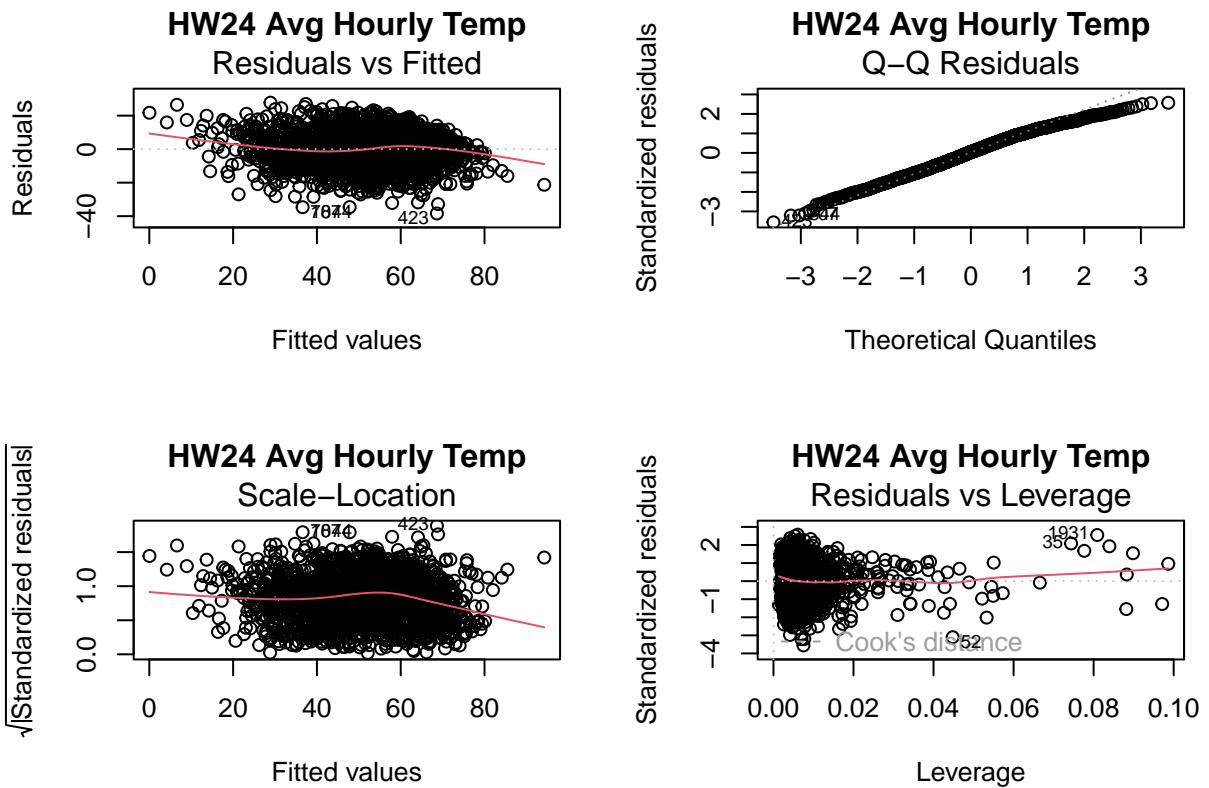
```
#HW24_S02
mod = lm(HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="S02 Max Avg 5 min")
```



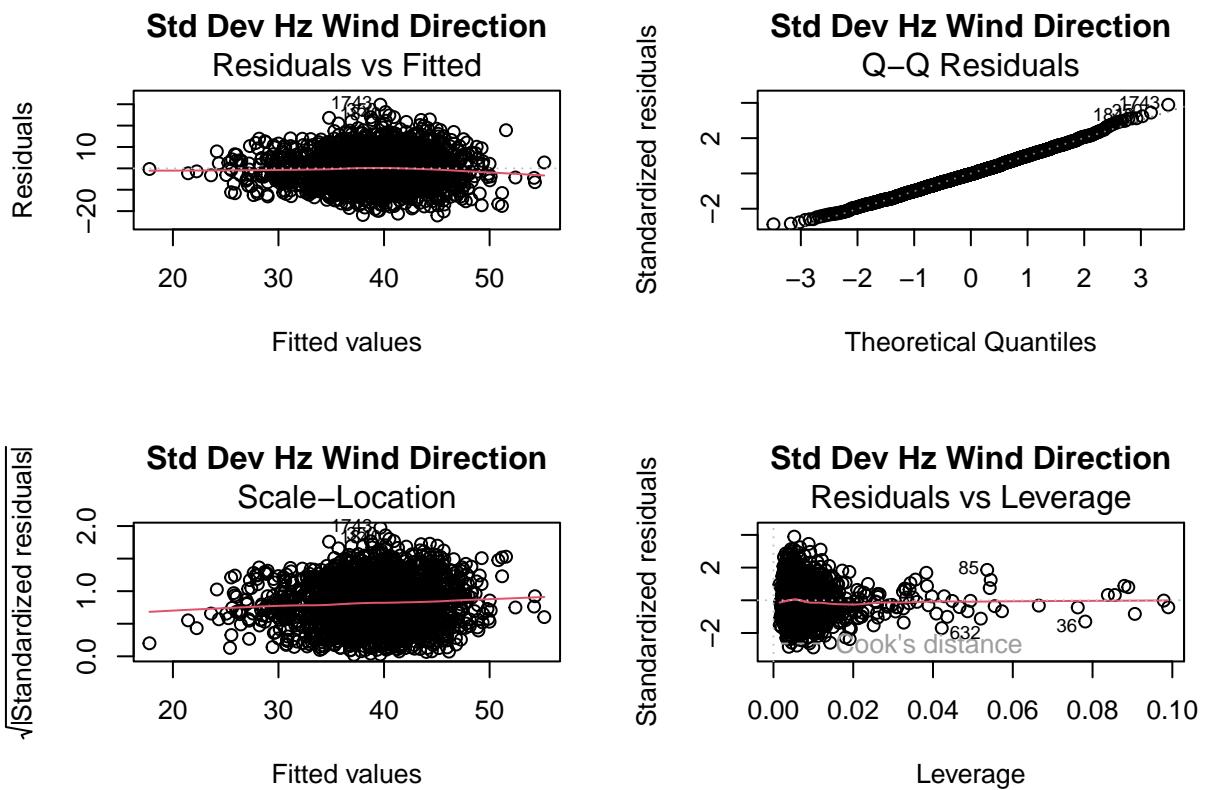
```
#HW24 Relative Humidity
mod = lm(HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean ~ .,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="Relative Humidity")
```



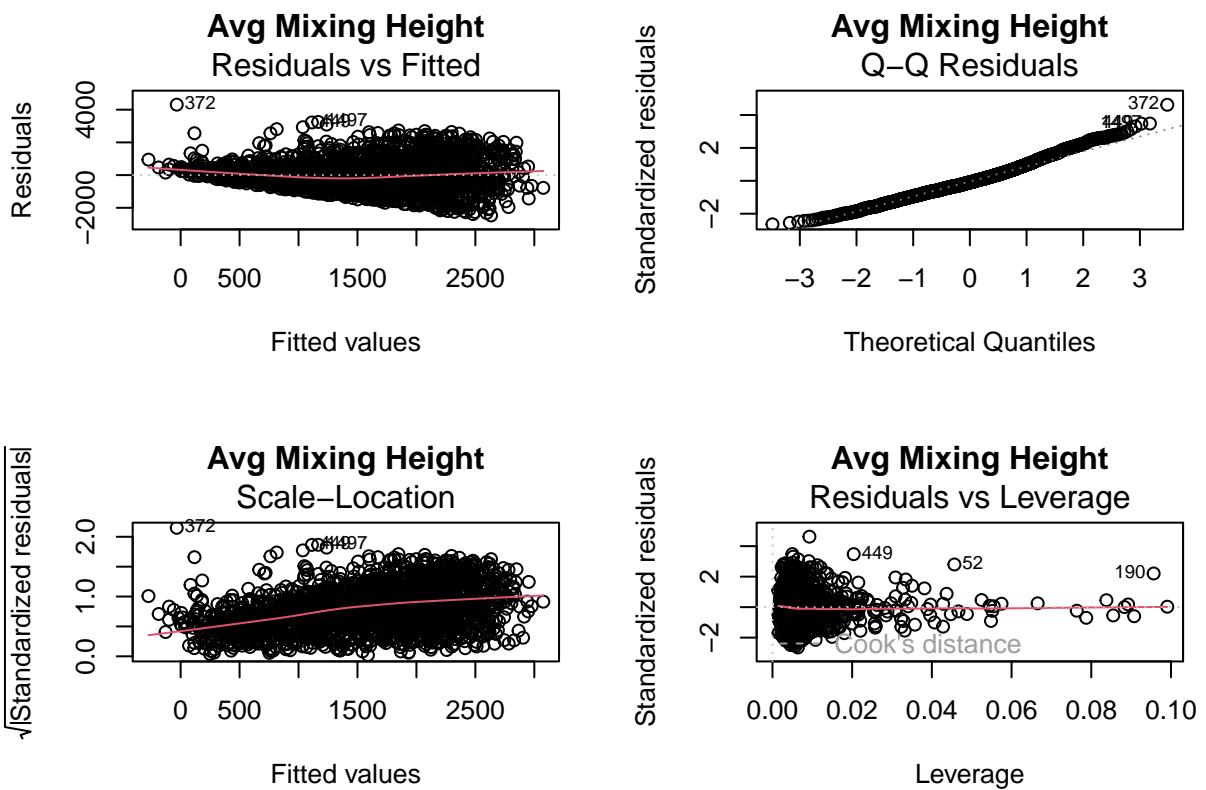
```
#HW24_Temp
mod = lm(HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean ~ .,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="HW24 Avg Hourly Temp")
```



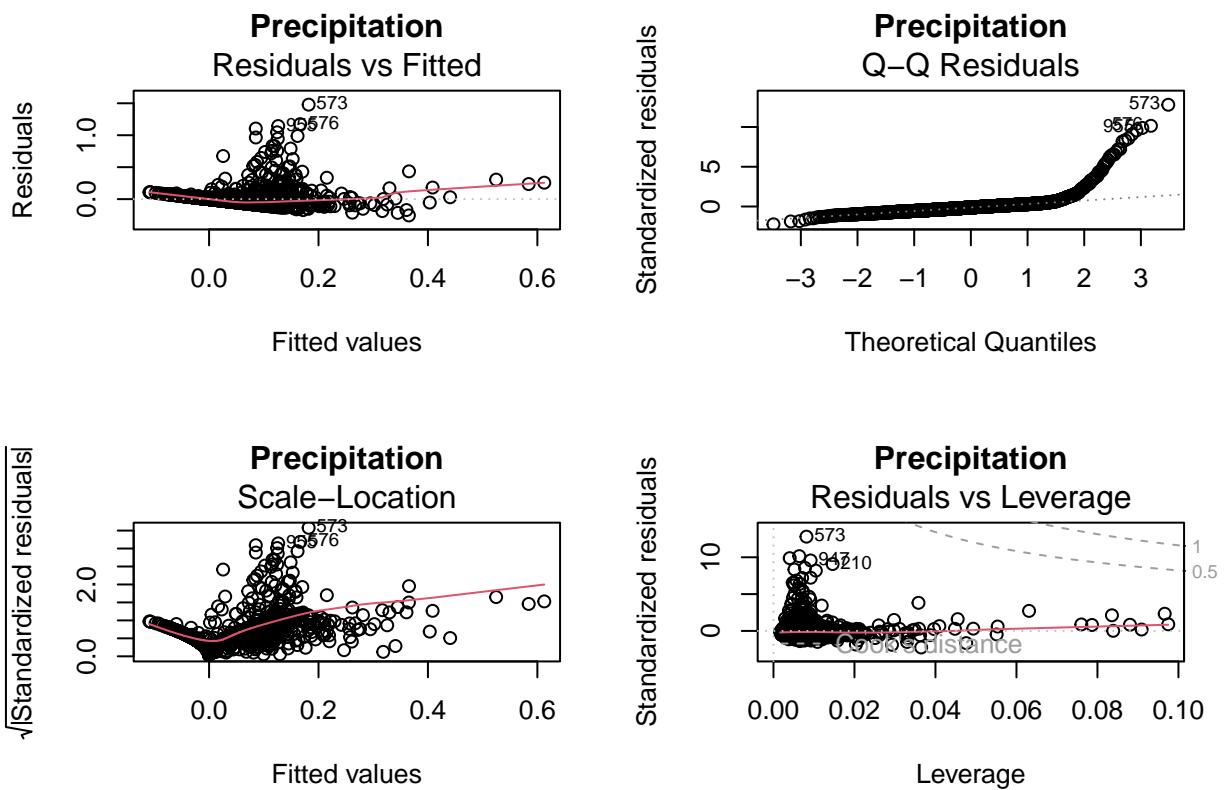
```
#HW24 Std Dev Hz Wind Direction
mod = lm(HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean~.,
         data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="Std Dev Hz Wind Direction")
```



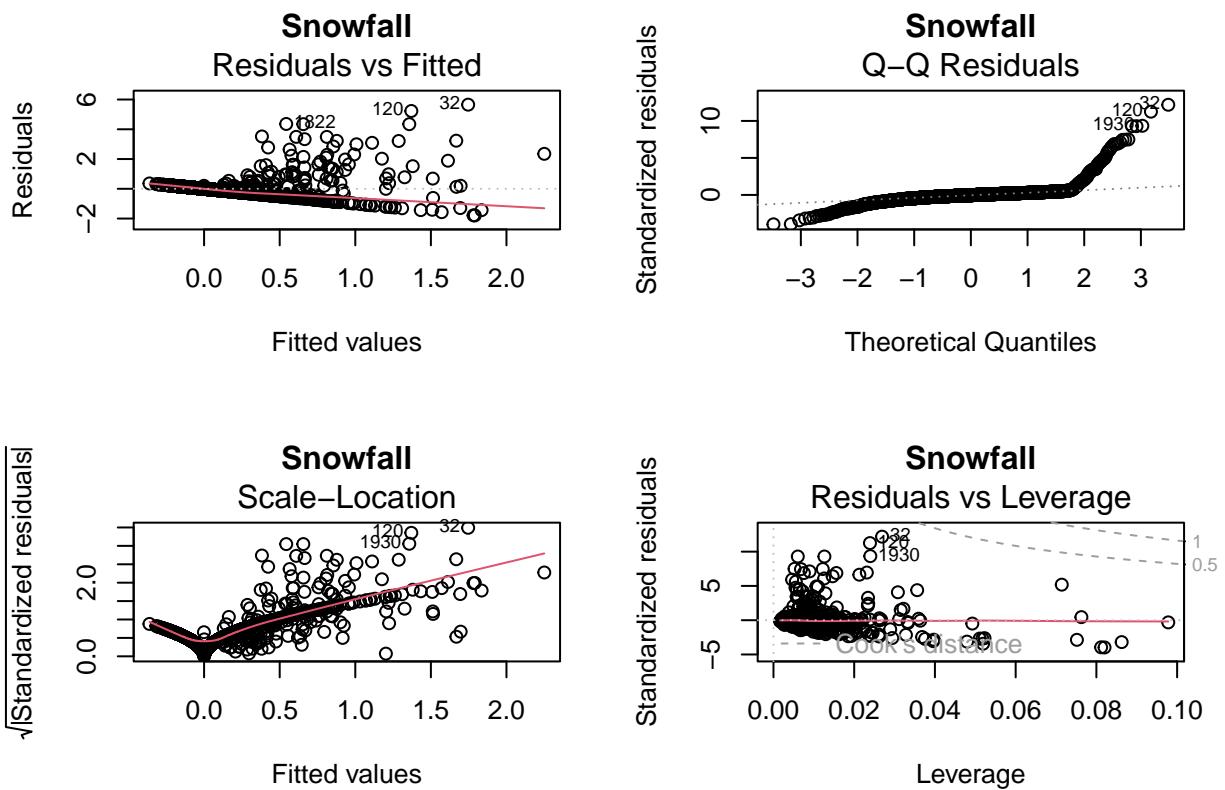
```
#IGRA Avg Mixing Height
mod = lm(Average.Mixing.Height~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="Avg Mixing Height")
```



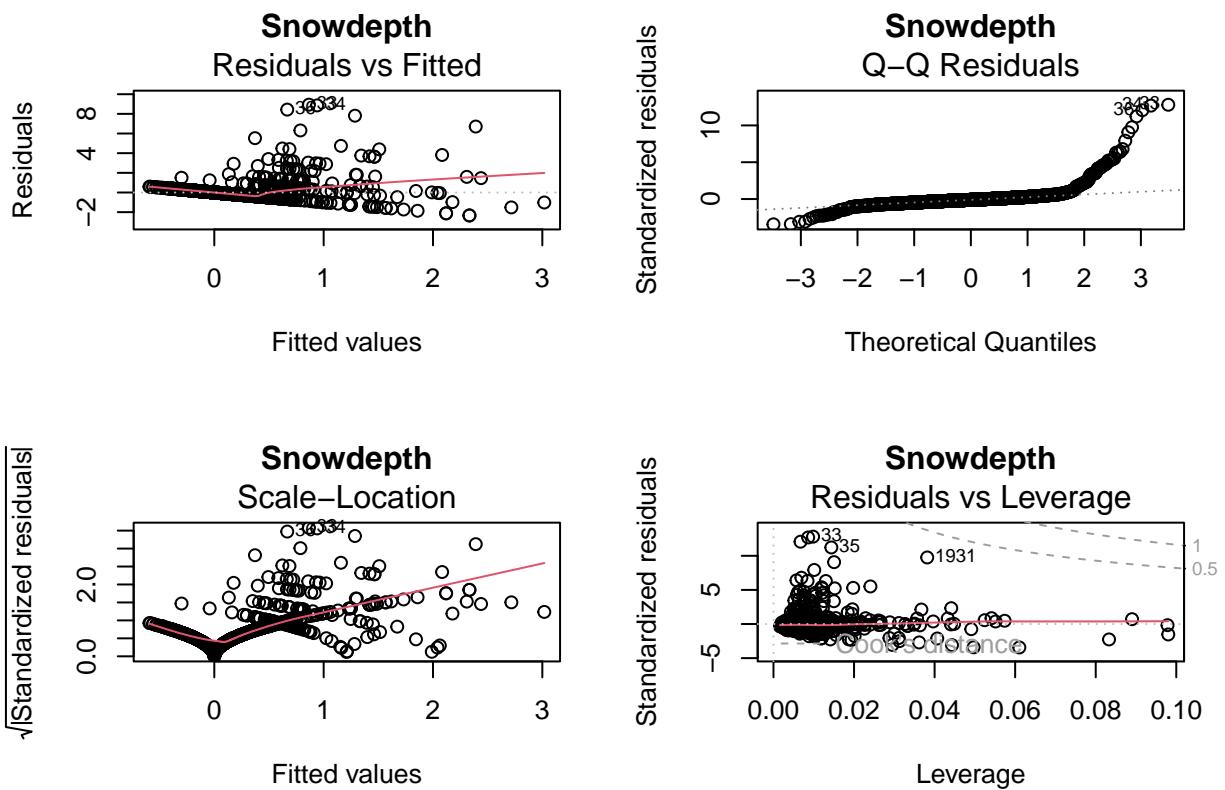
```
#COS Precipitation
mod = lm(Precipitation~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="Precipitation")
```



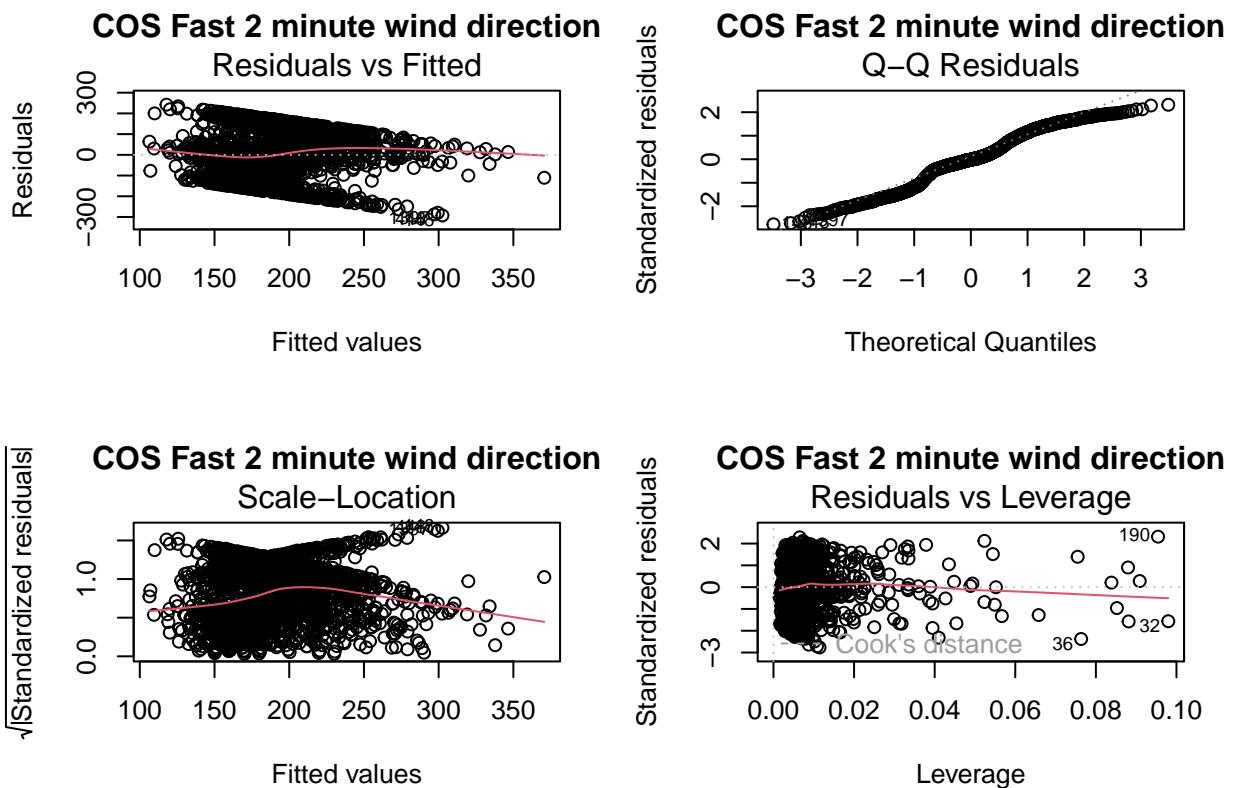
```
#COS Snowfall
mod = lm(Snowfall~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="Snowfall")
```



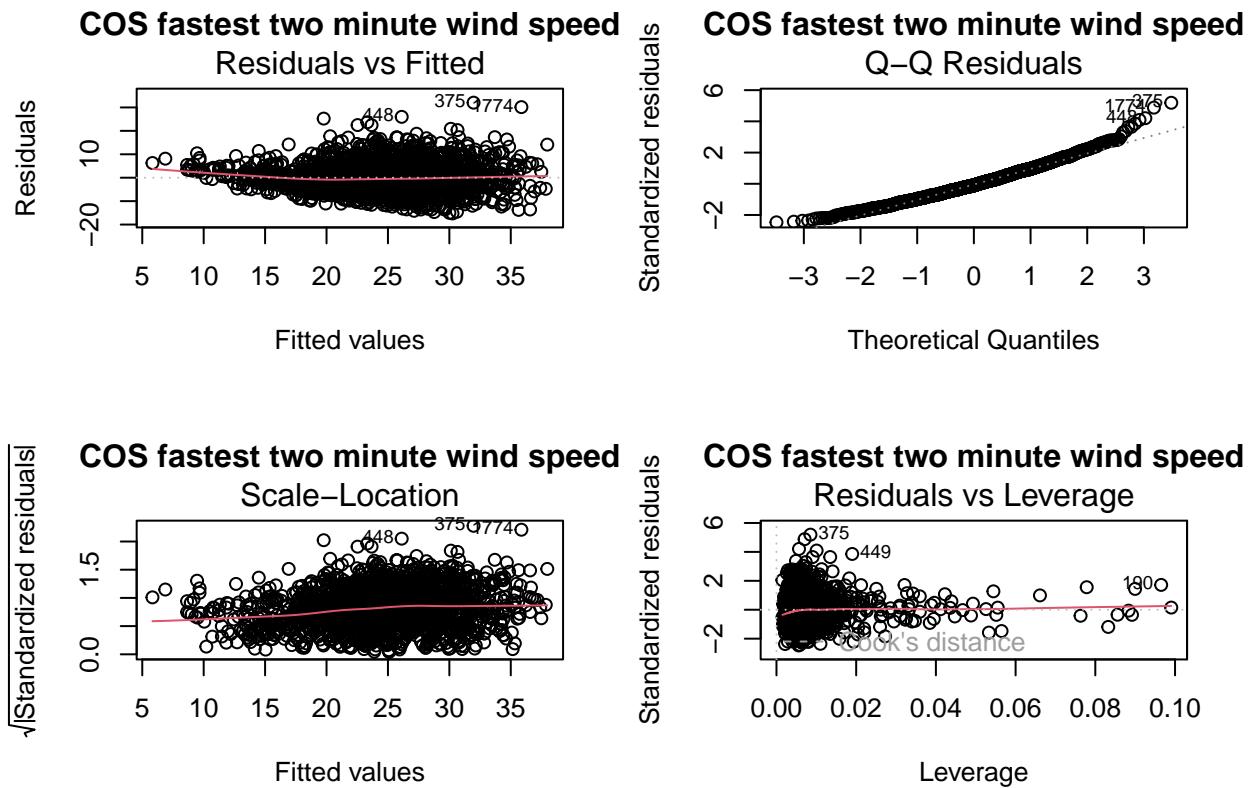
```
#COS Snowdepth
mod = lm(Snow.depth~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="Snowdepth")
```



```
#COS fastest two minute wind direction
mod = lm(Direction.of.fastest.2.minute.wind~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="COS Fast 2 minute wind direction")
```

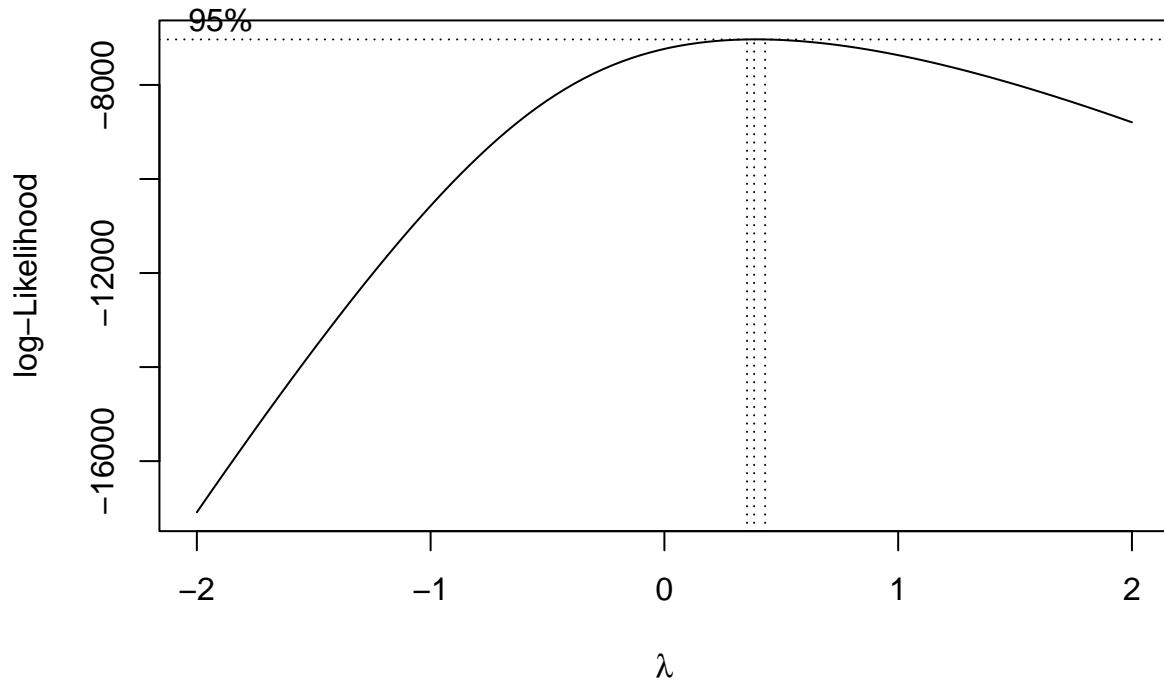


```
#COS fastest two minute wind speed
mod = lm(Fastest.2.minute.wind.speed~.,
          data = ozone_8_hr)
par(mfrow = c(2, 2))
plot(mod, main="COS fastest two minute wind speed")
```



#Transform Mixing Height to meet linear model assumptions

```
#transform variable with box cox
ozone_transformed = ozone_8_hr
pre_T_fit = lm(Average.Mixing.Height ~ ., data = ozone_transformed)
#select transformation of response
boxcox(pre_T_fit)
lam <- boxcox(pre_T_fit)$x[which.max(boxcox(pre_T_fit)$y)] # identify response transform
```

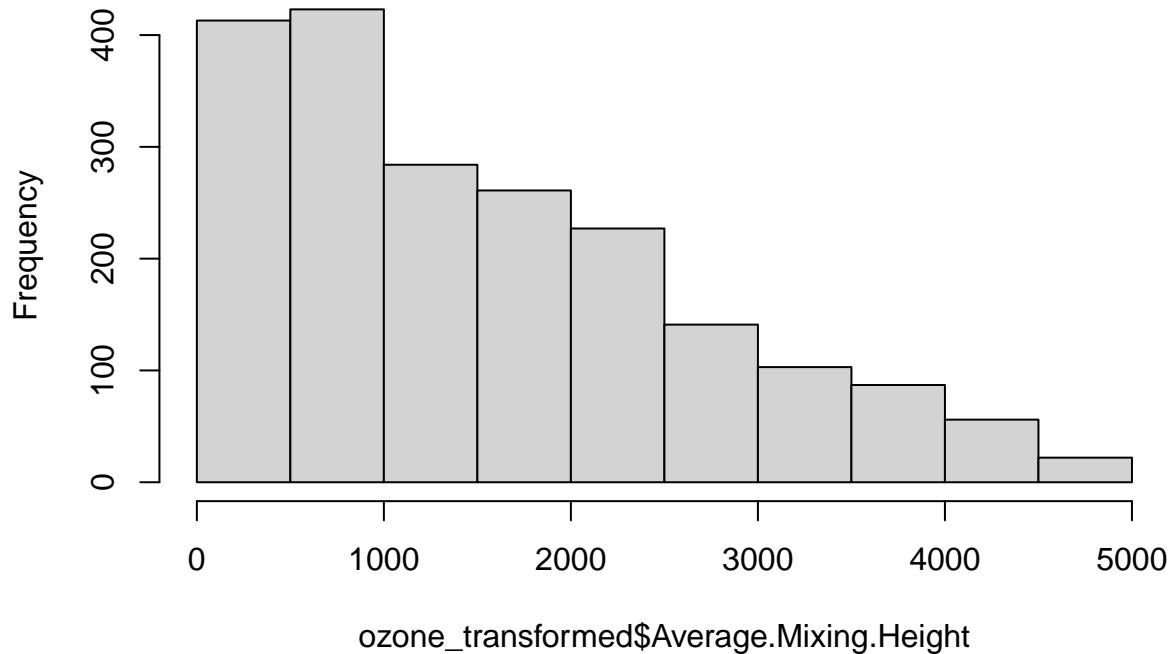


```
lam
```

```
## [1] 0.3838384
```

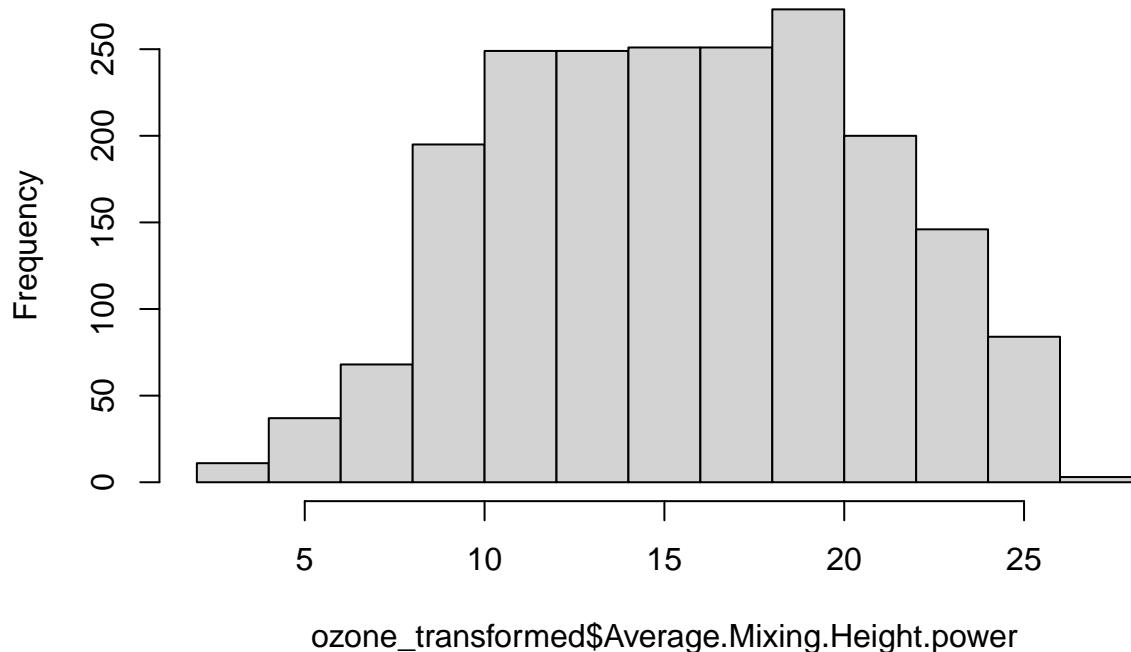
```
#transform response and store
ozone_transformed$Average.Mixing.Height.power <- ozone_transformed$Average.Mixing.Height^(lam)
#examine histograms of variable before and after transformation
hist(ozone_transformed$Average.Mixing.Height)
```

Histogram of ozone_transformed\$Average.Mixing.Height



```
hist(ozone_transformed$Average.Mixing.Height.power)
```

Histogram of ozone_transformed\$Average.Mixing.Height.power



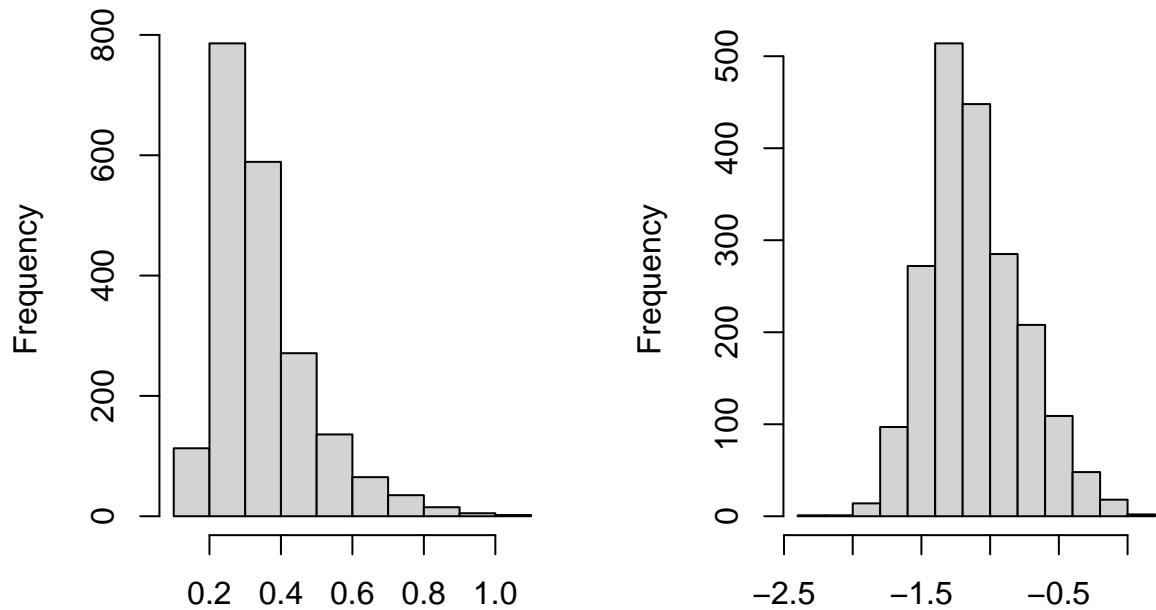
Transform additional variables for linear model

```
#log transform
#delete non-transformed columns
ozone_transformed <- ozone_transformed %>%
  mutate(log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean = log(HW24_Carbon.monoxide_8.HR.
    log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean = log(HW24_SO2.max.5.min.avg_1.HOUR_Arithmet

#A -Inf value is produced in log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean, so manually replace with
log_sulfur <- ozone_transformed$log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean
for (ii in 1:length(log_sulfur)) {
  if (log_sulfur[ii] == -Inf) {
    ozone_transformed$log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean[ii] = -10
  }
}

#examine histograms of variable before and after transformation
par(mfrow=c(1,2))
hist(ozone_8_hr$HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean)
hist(ozone_transformed$log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean)
```

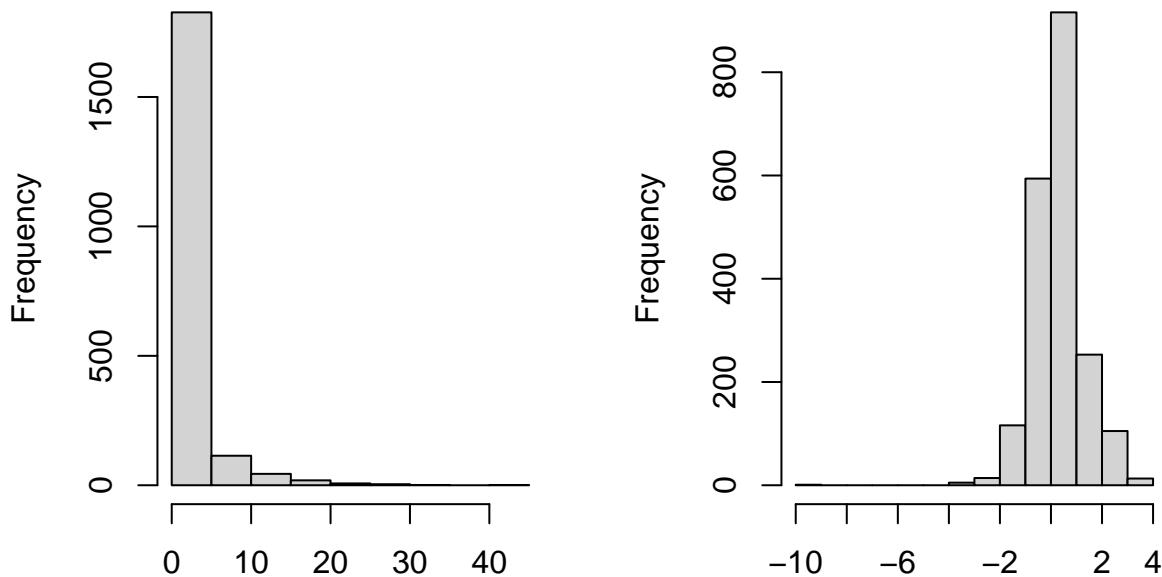
N24_Carbon.monoxide_8.HR.RUN.Avg_HW24_Carbon.monoxide_8.HR.



4_Carbon.monoxide_8.HR.RUN.AVG.END.F_HW24_Carbon.monoxide_8.HR.RUN.AVG.I

```
par(mfrow=c(1,2))
hist(ozone_8_hr$HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean) # log of Sulfur Dioxide creates too many
hist(ozone_transformed$log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean)
```

_8_hr\$HW24_SO2.max.5.min.avg_1formed\$log_HW24_SO2.max.5.min.



_hr\$HW24_SO2.max.5.min.avg_1.HOUR_Amed\$log_HW24_SO2.max.5.min.avg_1.HOI

```
#drop non-transformed variables
```

```
dim(ozone_8_hr)
```

```
## [1] 2017 15
```

```
ozone_transformed <- ozone_transformed %>% dplyr::select(-HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetric.Mean, -Precipitation, -Snowfall, -Snow.depth) #drop unneeded columns
```

```
## [1] "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetric.Mean"  
## [2] "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetric.Mean"  
## [3] "HW24_Wind.Direction...Resultant_1.HOUR_Arithmetric.Mean"  
## [4] "HW24_Relative.Humidity_1.HOUR_Arithmetric.Mean"  
## [5] "HW24_Outdoor.Temperature_1.HOUR_Arithmetric.Mean"  
## [6] "HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetric.Mean"  
## [7] "Direction.of.fastest.2.minute.wind"  
## [8] "Fastest.2.minute.wind.speed"  
## [9] "Date"  
## [10] "Average.Mixing.Height.power"  
## [11] "log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetric.Mean"  
## [12] "log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetric.Mean"
```

```

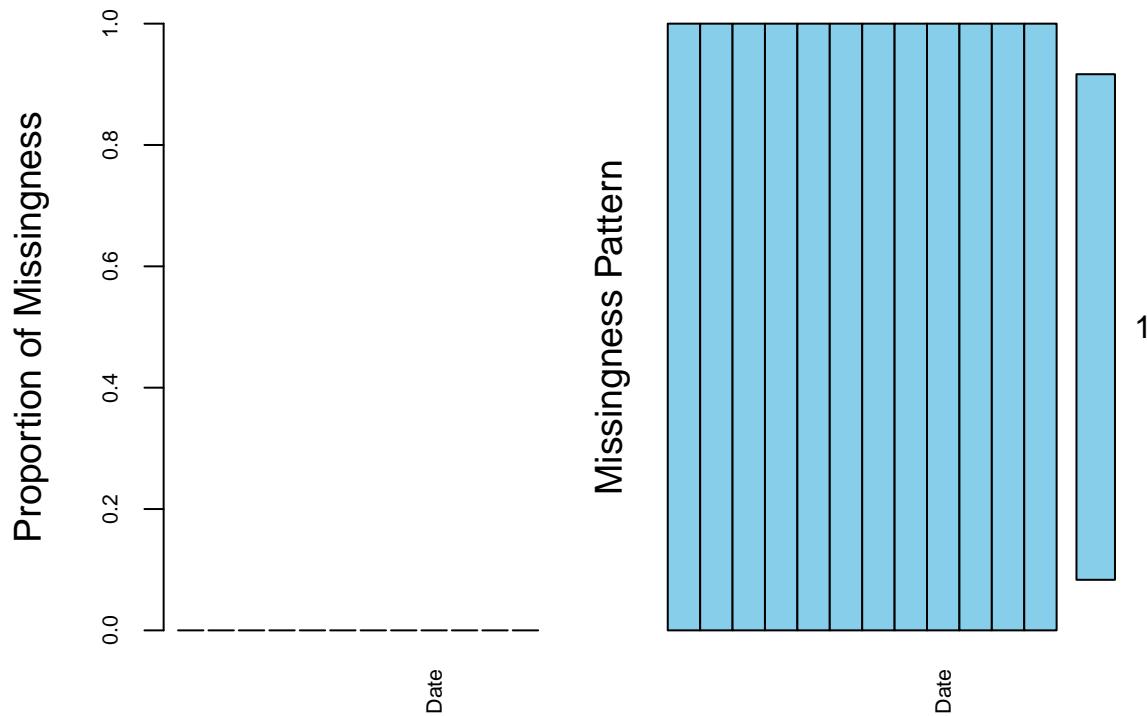
dim(ozone_transformed)

## [1] 2017    12

#check transformed df for missingness

aggr(ozone_transformed, numbers=TRUE, sortVars=TRUE, cex.axis=.7, gap=3,
      ylab=c("Proportion of Missingness","Missingness Pattern"))

```

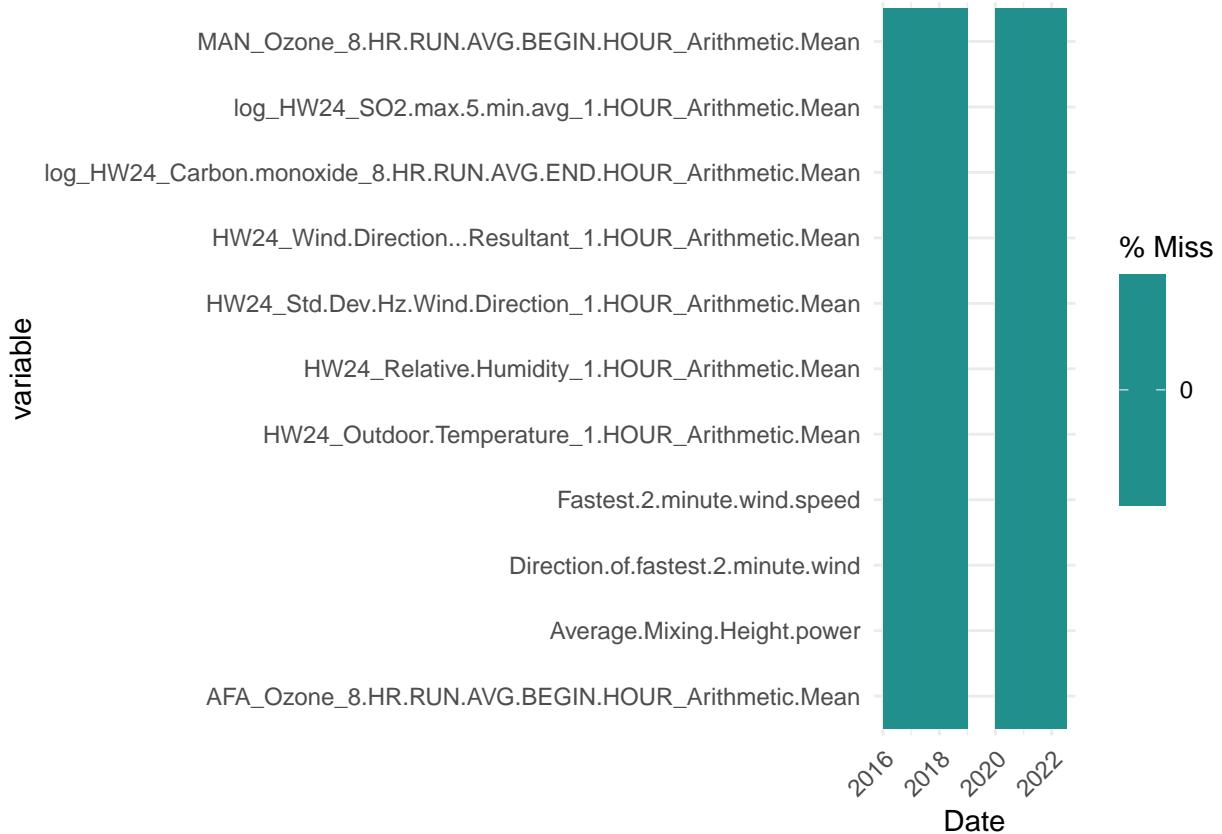


```

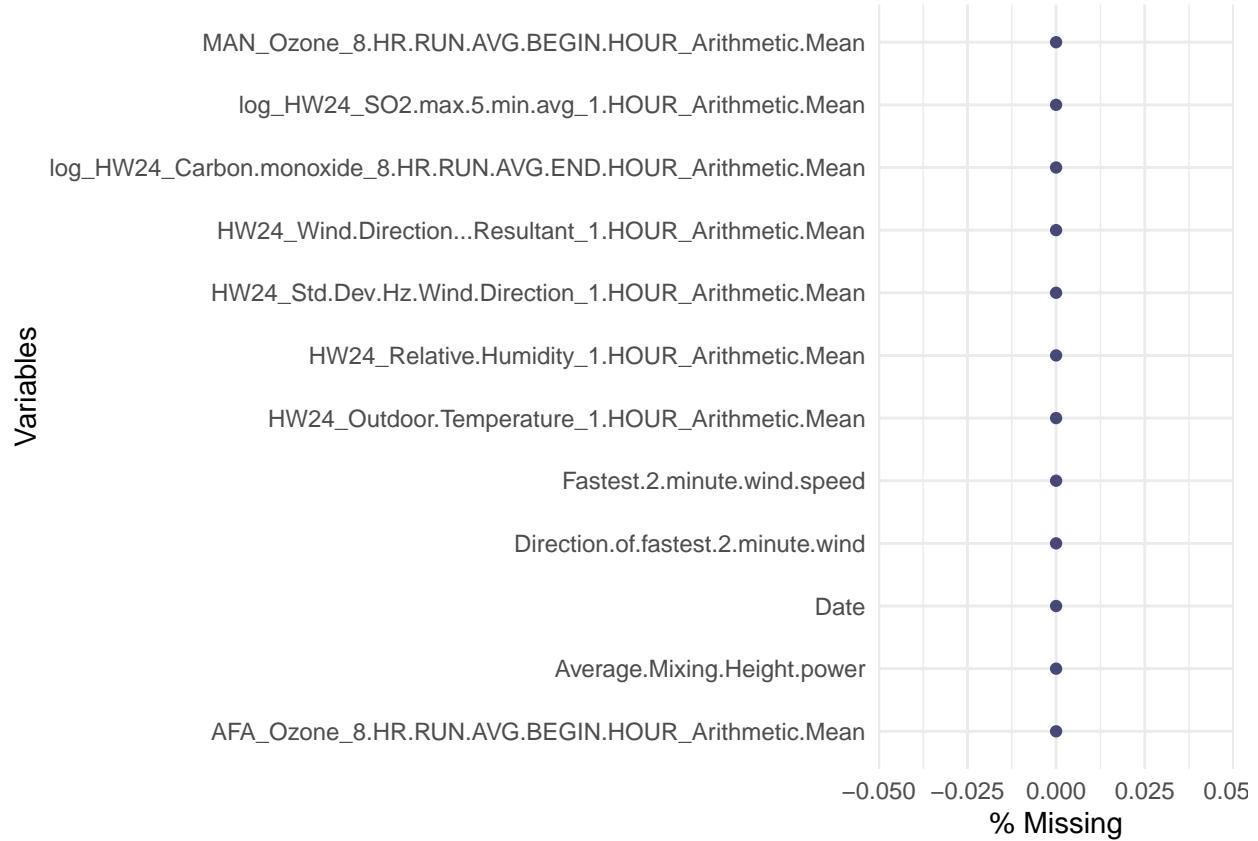
##
##  Variables sorted by number of missings:
##                                         Variable Count
##          AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean      0
##          MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean      0
##          HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean      0
##          HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean      0
##          HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean      0
##          HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean      0
##          Direction.of.fastest.2.minute.wind      0
##          Fastest.2.minute.wind.speed      0
##          Date      0
##          Average.Mixing.Height.power      0
## log_HW24_Carbon.monoxide_8.HR.RUN.AVG-END.HOUR_Arithmetic.Mean      0
## log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean      0

```

```
gg_miss_fct(ozone_transformed, Date)
```



```
gg_miss_var(ozone_transformed, show_pct = TRUE)
```



MODELING WITH CARET

#Hyper-parameter tuning

```
#Tuning Functions
#LASSO
ozone_tuning_LASSO <- function(dataused, ozone_col,lambdaLASSO){
  set.seed(28)
  training = trainControl(method = "cv", number = 5) #5 for computational feasibility
  #fit linear regression with LASSO
  fit_LASSO = train(formula(paste(ozone_col, " ~ .")),
    data = dataused,
    method = "glmnet",
    trControl = training,
    tuneGrid = expand.grid(alpha=c(1),lambda=lambdaLASSO))
  print("LASSO Model has been fitted")#debug
  lam = paste("Best lambda for LASSO is:", fit_LASSO$bestTune$lambda )
  print(lam)
  print(Sys.time())
  return(fit_LASSO)
}
#Boosted Trees
ozone_tuning_BOOST <- function(dataused, ozone_col,lambdaBOOST){
  set.seed(28)
```

```

training = trainControl(method = "cv", number = 5) #5 for computational feasibility
#fit Boosted Trees
fit_boost = train(formula(paste(ozone_col, " ~ .")),
                  data = dataused,
                  method = "gbm",
                  trControl = training,
                  tuneGrid = expand.grid(interaction.depth = (1:3), #use max interaction dept
                                         n.trees = 300, #use 300 trees (more trees is ideal, but comp
                                         shrinkage = lambdaBOOST, #tuned hyperparameter vector
                                         n.minobsinnode = 10), # use default
                  verbose=FALSE)
print("Boosted Trees has been fitted")
lam = paste("Best lambda for Boosted Trees is:", fit_boost$bestTune$shrinkage)
print(lam)
print(Sys.time())
return(fit_boost)
}

#Artificial Neural Net
ozone_tuning_ANN <- function(dataused, ozone_col,lambdaANN){
  set.seed(28)
  training = trainControl(method = "cv", number = 5) #5 for computational feasibility
  #fit Artificial Neural Net
  fit_ANN = train(formula(paste(ozone_col, " ~ .")),
                  data = dataused,
                  method = "nnet",
                  tuneGrid = expand.grid(size = (1:3), decay = lambdaANN), #limited for computational .
                  trace = FALSE,
                  preProc = c("center", "scale"), #good idea to prevent one predictor from overwhelming
                  trControl = training,
                  verbose=FALSE)
  print("Artificial Neural Net has been fitted") #debug
  lam = paste("Best lambda for Artificial Neural Net is:", fit_ANN$bestTune$decay)
  print(lam)
  print(Sys.time())
  return(fit_ANN)
}

#run all models
ozone_tune_models <- function(ozone_data_frame, ozone_col,lambdaLASSO, lambdaBOOST, lamdaANN){
  set.seed(28)
  #lasso
  lasso = ozone_tuning_LASSO(ozone_data_frame,ozone_col,lambdaLASSO)
  #store plot
  l <- gf_line(RMSE ~ lambda, data = lasso$results) %>%
    gf_refine(coord_cartesian(xlim = c(-.001, .005), ylim = c(0, .015))) %>%
    gf_vline(xintercept =~ lasso$bestTune$lambda,
              color = "red")
  #boost
  boost = ozone_tuning_BOOST(ozone_data_frame,ozone_col,lambdaBOOST)
  b <- plot(boost) #store plot
  #ann
  ann = ozone_tuning_ANN(ozone_data_frame,ozone_col, lamdaANN)
  a <- plot(ann) #store plot
}

```

```

print(1)
print(b)
print(a)
best_lambdas = c(lasso$bestTune$lambda, boost$bestTune$shrinkage, ann$bestTune$decay)
print(best_lambdas)
return(best_lambdas)
}

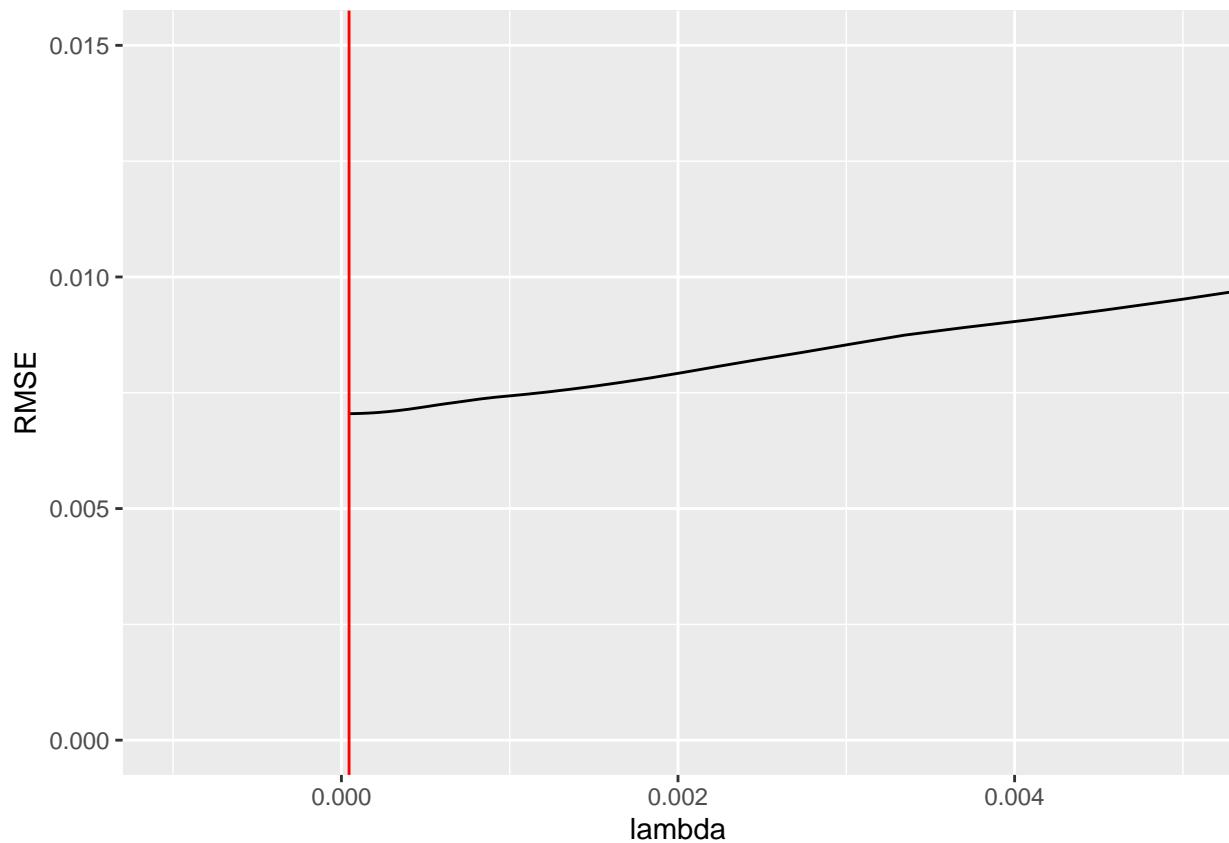
#AFA tuning #1

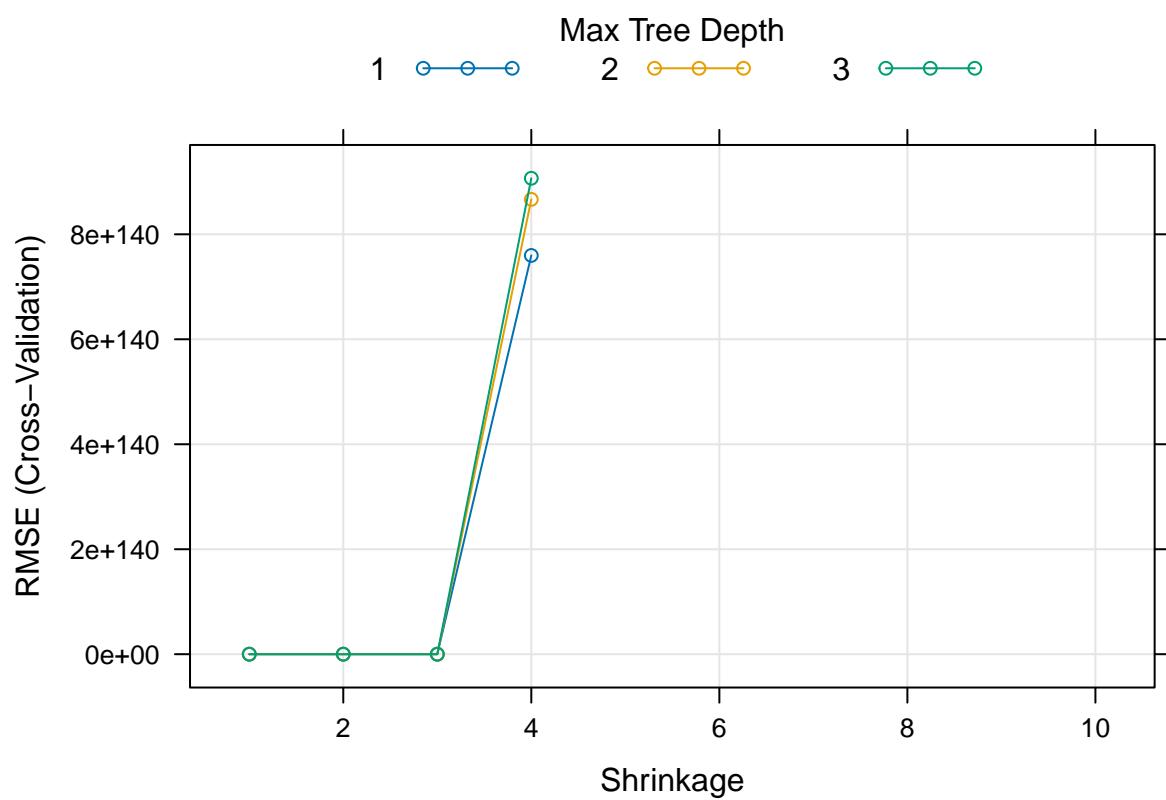
#dataset
afa_8_hr = ozone_8_hr[-c(2)]#use only AFA measurements
####Starting Values
#LASSO
lassoLambda_AFA = exp(-100:15/10) #initial search space
#GBM
boostLambda_AFA = seq(1,10,1) #examine 1-10 for initial search
#ANN
annLambda_AFA = seq(1,10,1) #examine 1-10 for initial search
##run models
afa_tune_1 = ozone_tune_models(afa_8_hr, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",
                                lassoLambda_AFA, boostLambda_AFA, annLambda_AFA)

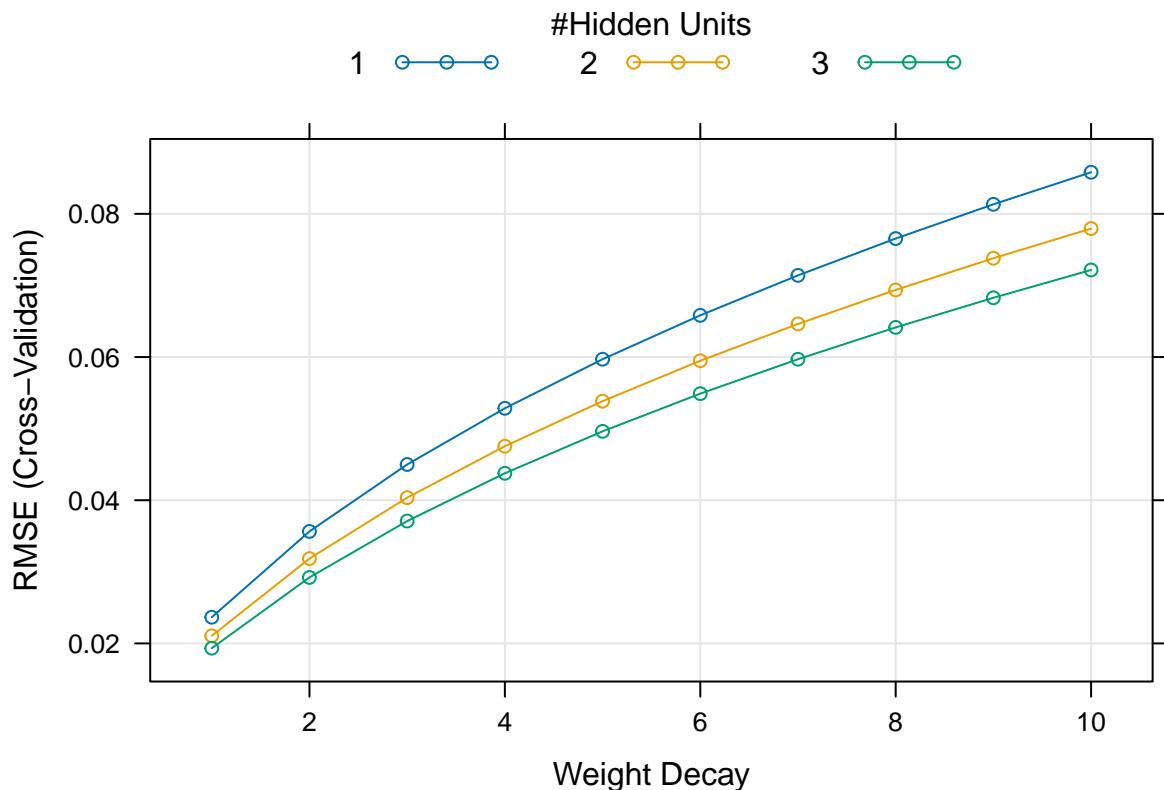
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 4.53999297624849e-05"
## [1] "2023-12-09 23:06:58 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 1"
## [1] "2023-12-09 23:07:27 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 1"
## [1] "2023-12-09 23:07:32 MST"

```







```

## [1] 4.539993e-05 1.000000e+00 1.000000e+00

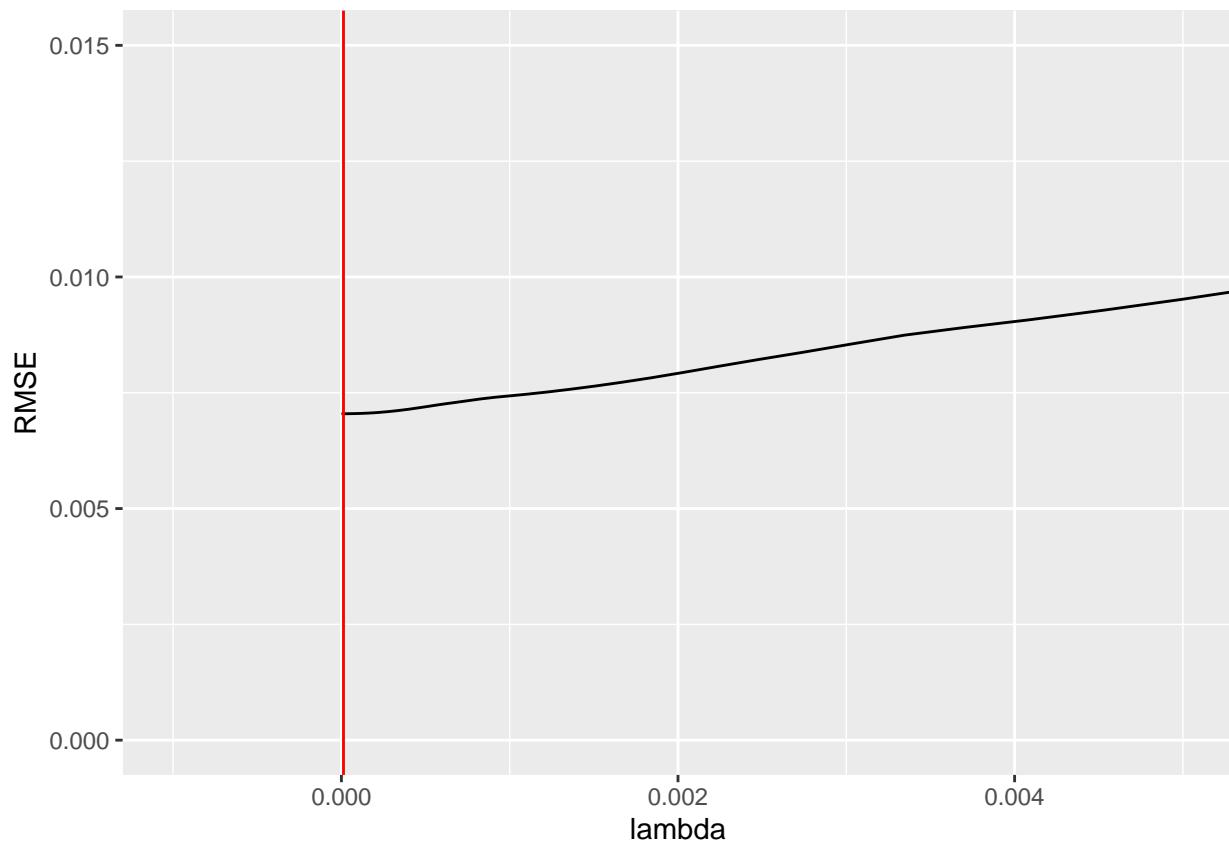
#AFA tuning #2

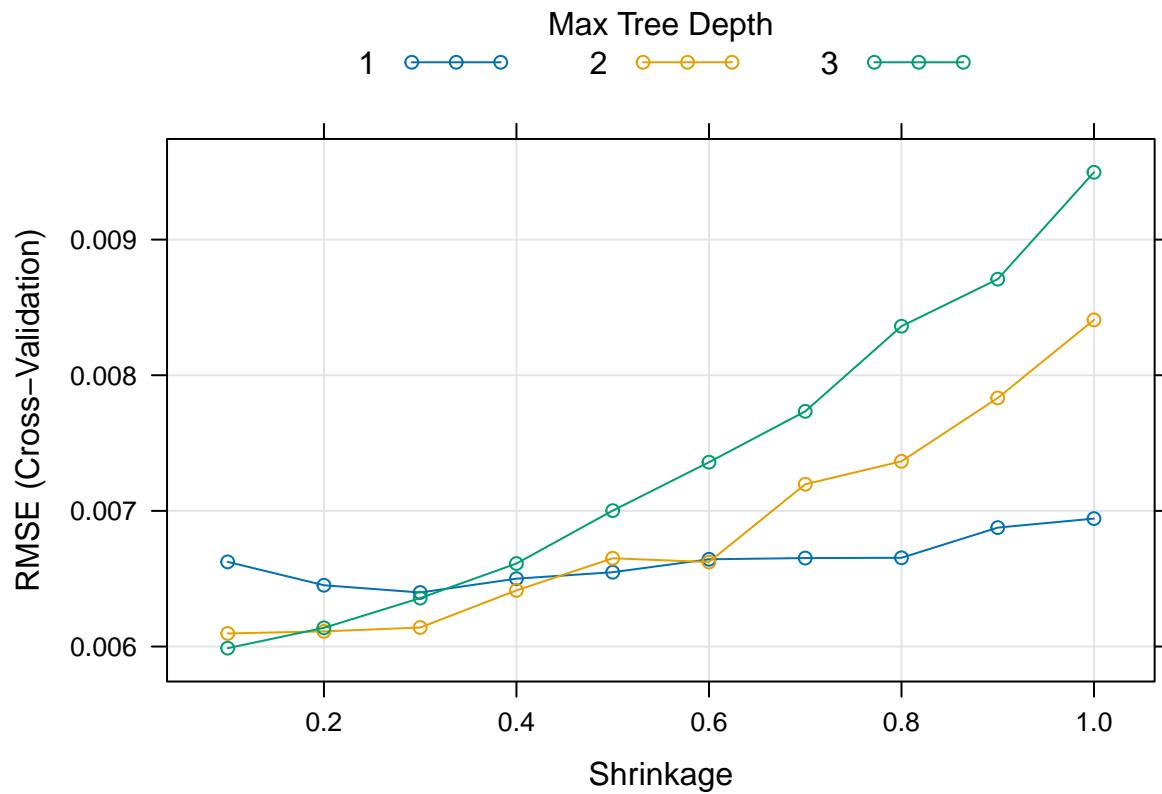
####Updated Values
lassoLambda_AFA = exp(-150:-35/10) #returned to get smaller values
boostLambda_AFA = seq(.1,afa_tune_1[2],.1) #best tune is 1, lower bound. Examine 10x finer scale
annLambda_AFA = seq(.1,afa_tune_1[3],.1) #best tune is 1, lower bound. Examine 10x finer scale
##run models
afa_tune_2 = ozone_tune_models(afa_8_hr, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean", lassoLambda_AFA)

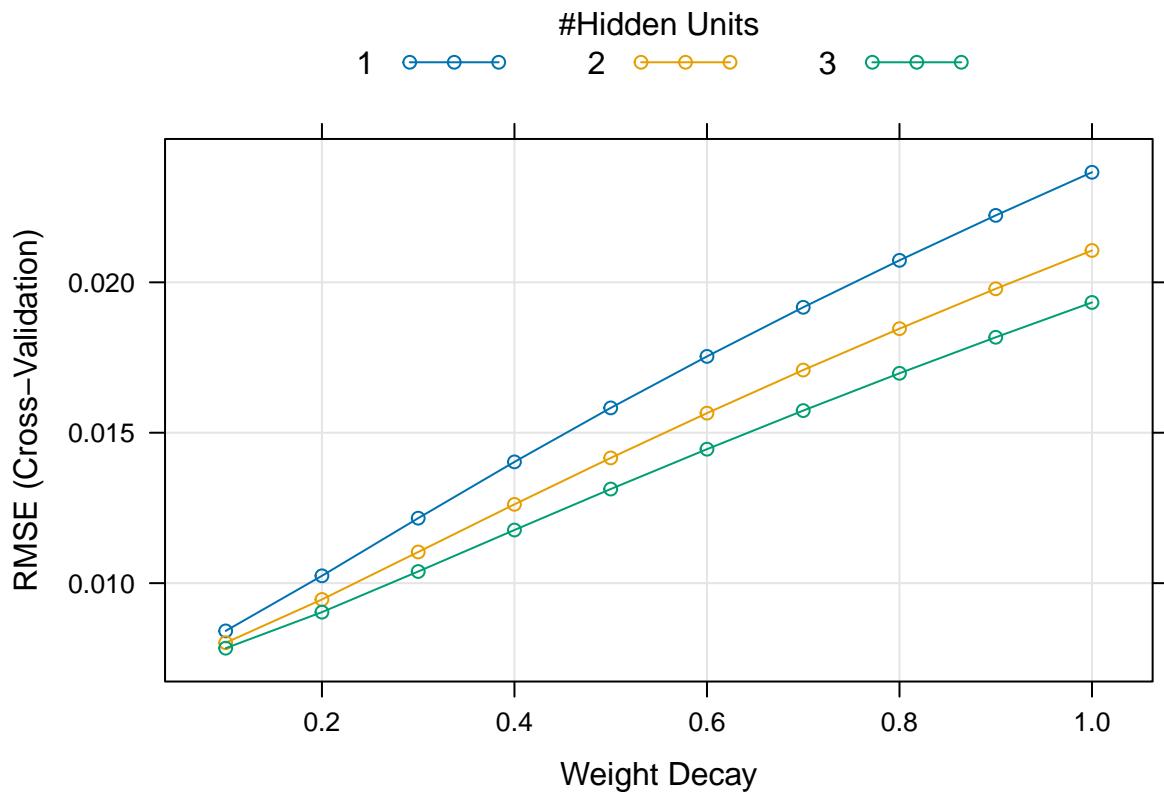
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.23729242617882e-05"
## [1] "2023-12-09 23:07:32 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.1"
## [1] "2023-12-09 23:08:01 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.1"
## [1] "2023-12-09 23:08:05 MST"

```







```

## [1] 1.237292e-05 1.000000e-01 1.000000e-01

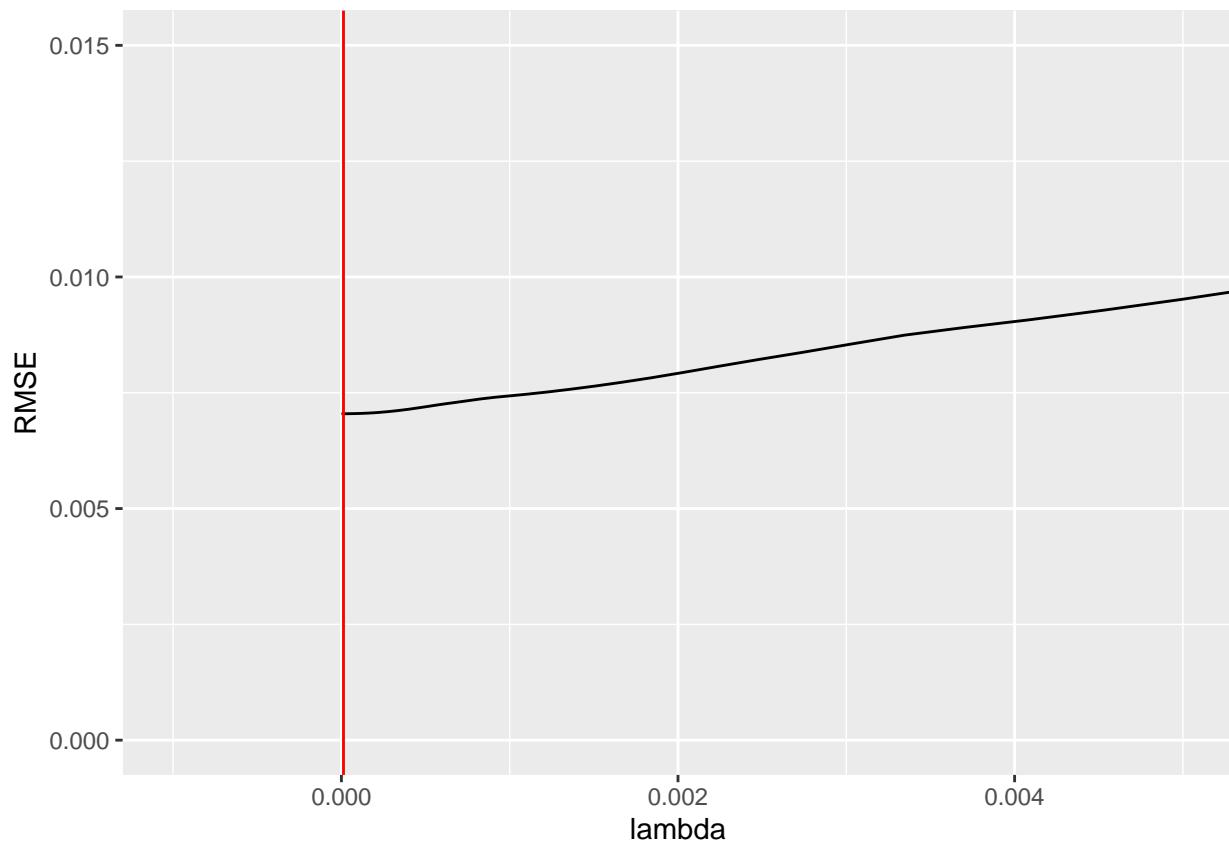
#AFA tuning #3

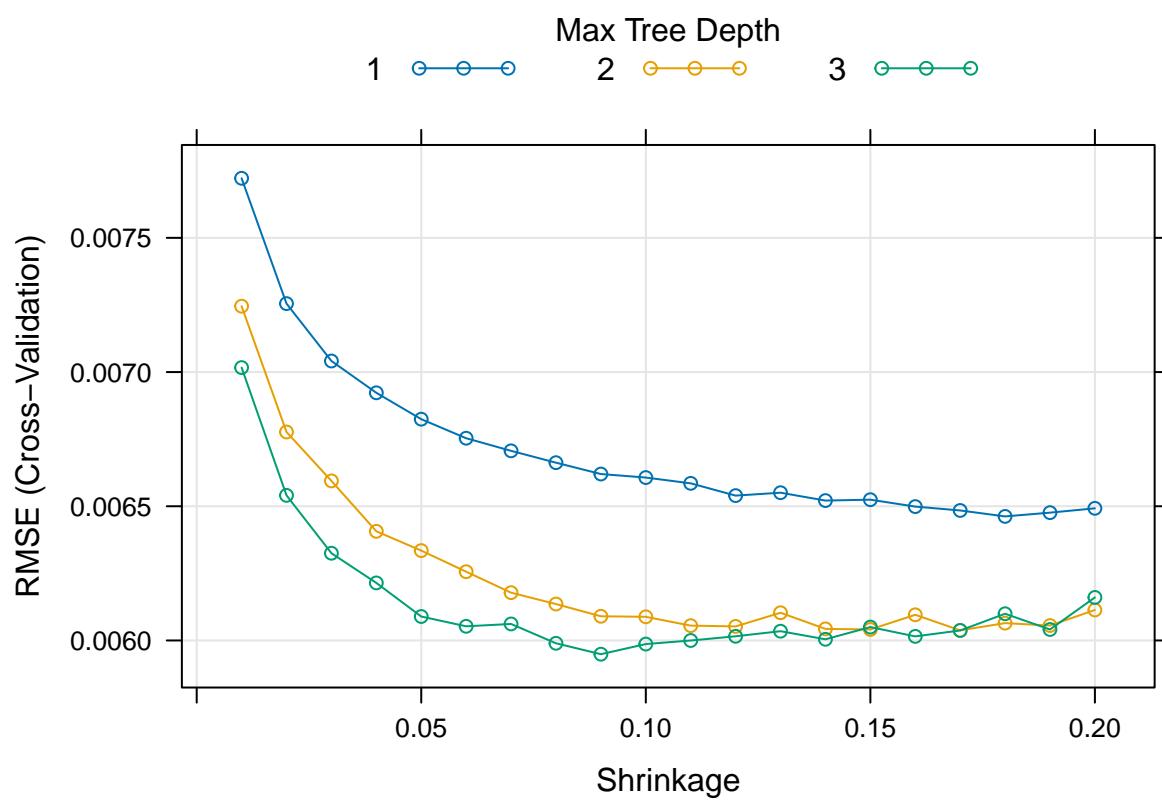
####Updated Values -
#Keep LASSO lambda since best tune lies within distribution of values
boostLambda_AFA = seq(.01,afa_tune_2[2]+.1,.01) #best tune is 0.1, lower & upper bound extended. Examine 10x finer
annLambda_AFA = seq(.01,afa_tune_2[3],.01) #best tune is 0.1, lower bound extended. Examine 10x finer
#run
afa_tune_3 = ozone_tune_models(afa_8_hr, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean", lassoLambda_AFA)

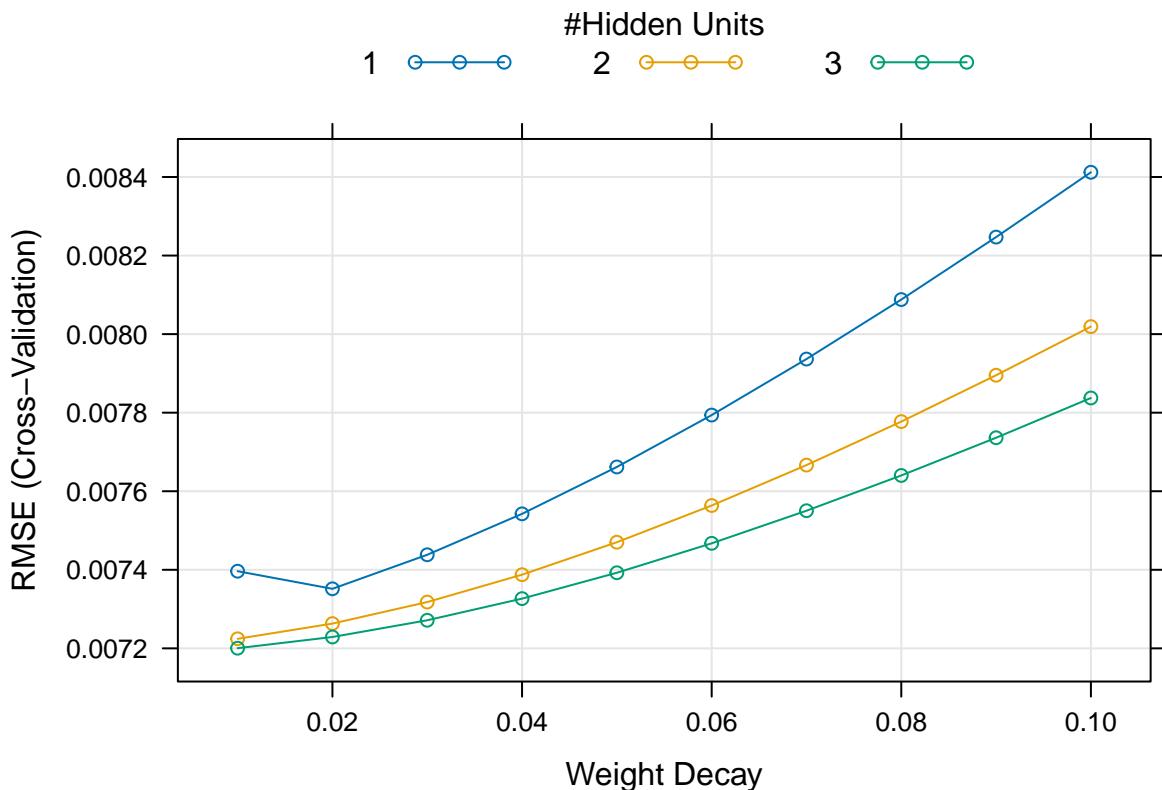
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.23729242617882e-05"
## [1] "2023-12-09 23:08:06 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.09"
## [1] "2023-12-09 23:09:02 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.01"
## [1] "2023-12-09 23:09:08 MST"

```







```

## [1] 1.237292e-05 9.000000e-02 1.000000e-02

#AFA tuning #4 (final Tune)

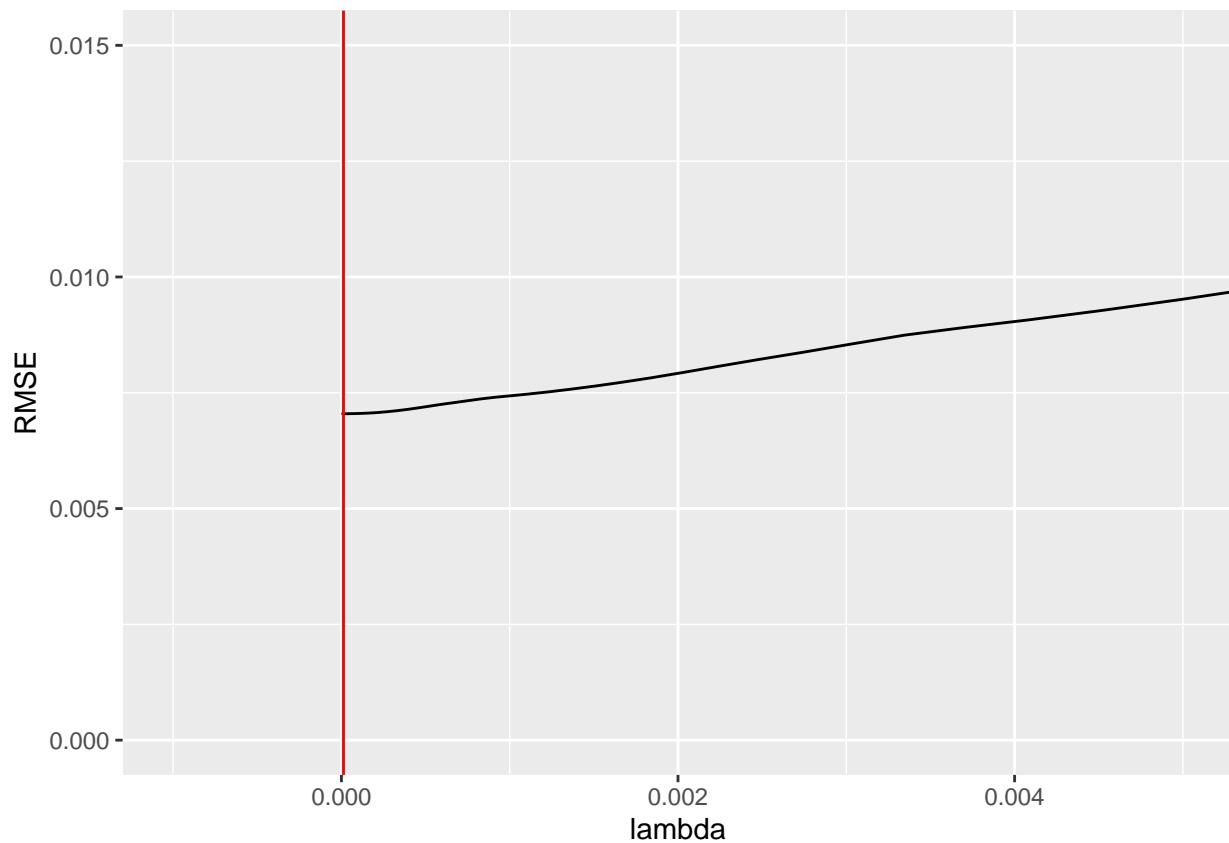
####Updated Values
boostLambda_AFA = seq(.05,afa_tune_2[2]+.05,.01) #shorten scale for final models
annLambda_AFA = seq(.001,.05,.001) #best tune is continuously smaller, lower bound extended. Examine 1

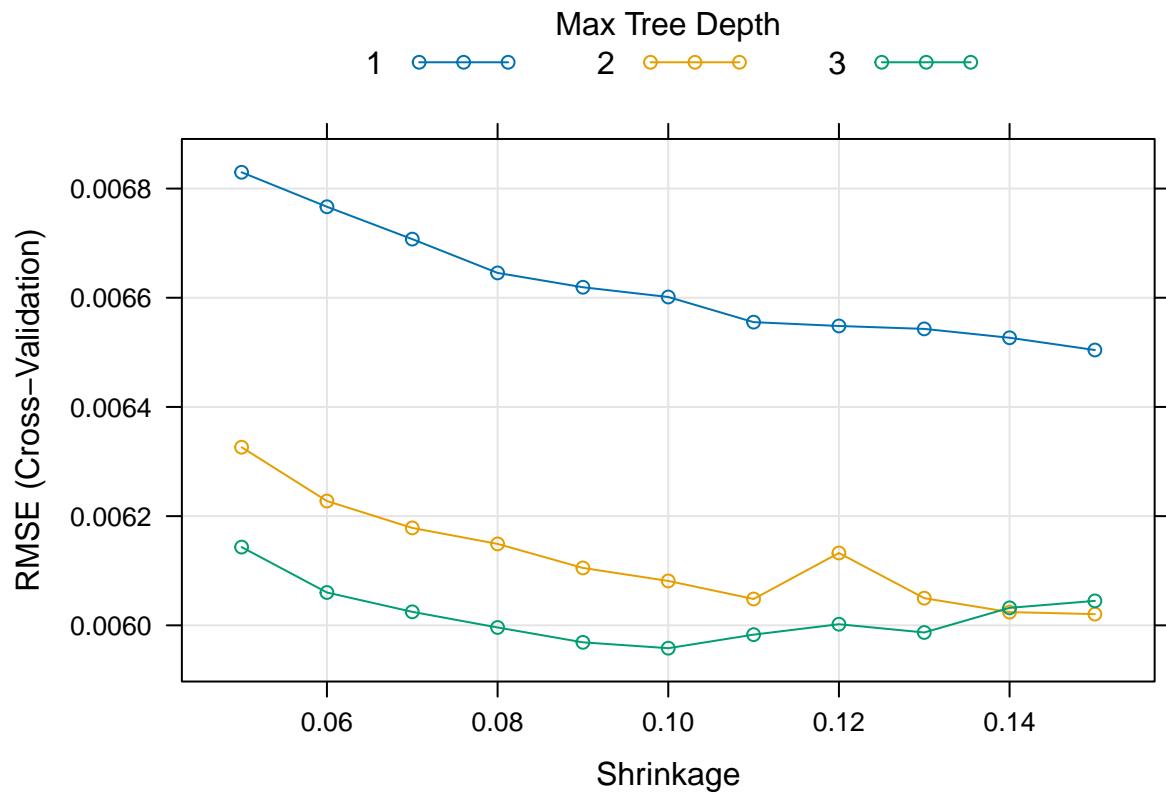
afa_tune_4 = ozone_tune_models(afa_8_hr, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean", lassoLambda_AFA)

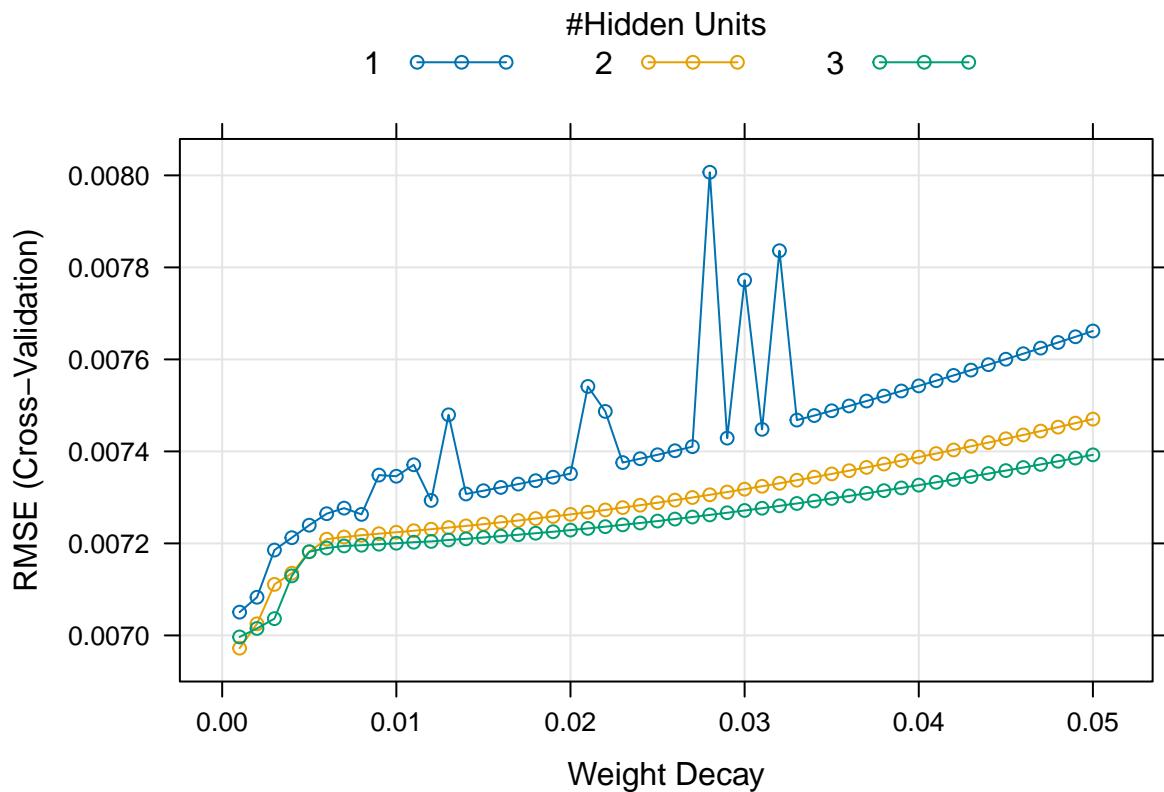
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.23729242617882e-05"
## [1] "2023-12-09 23:09:08 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.1"
## [1] "2023-12-09 23:09:39 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.001"
## [1] "2023-12-09 23:10:11 MST"

```





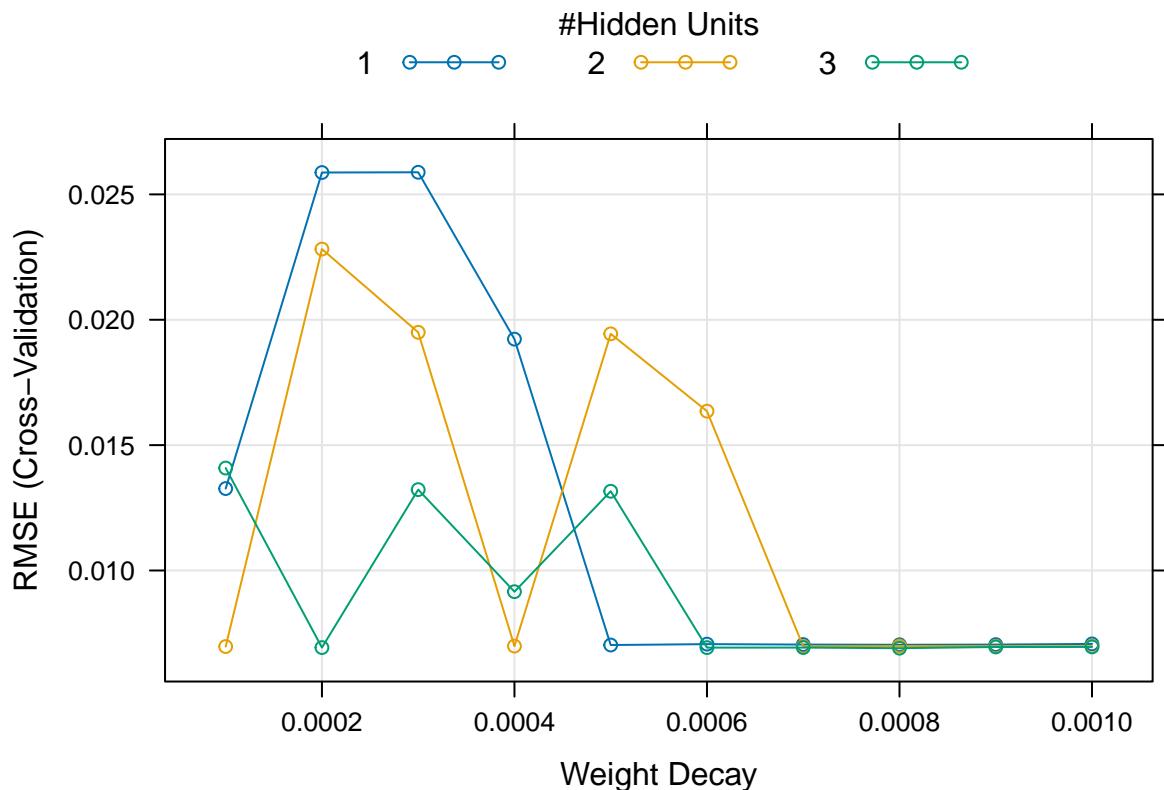


```
## [1] 1.237292e-05 1.000000e-01 1.000000e-03
```

```
#final ann finessing
annLambda_AFA = seq(.0001,.001,.0001) #best tune is continuously smaller, lower bound extended. Examining
ann = ozone_tuning_ANN(afa_8_hr, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",annLambda_AFA)

## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 8e-04"
## [1] "2023-12-09 23:10:19 MST"
```

```
plot(ann)
```



```
#AFA_transformed tuning #1
```

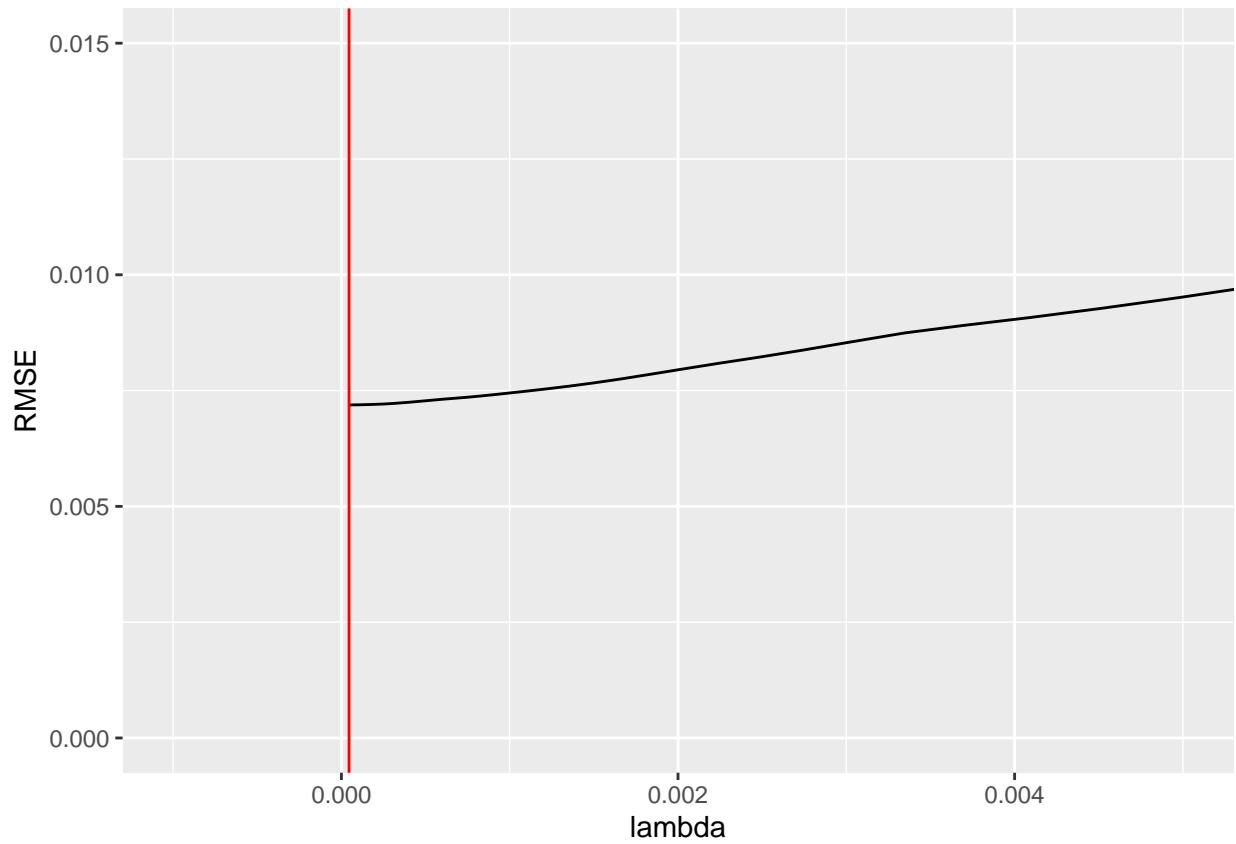
```
#dataset
afa_8_hr_trans = ozone_transformed[-c(2)] #use only AFA measurements
####Starting Values
#LASSO
lassoLambda_AFA_trans = exp(-100:15/10) #initial search space
#GBM
boostLambda_AFA_trans = seq(1,10,1) #examine 1-10 for initial search
#ANN
annLambda_AFA_trans = seq(1,10,1) #examine 1-10 for initial search

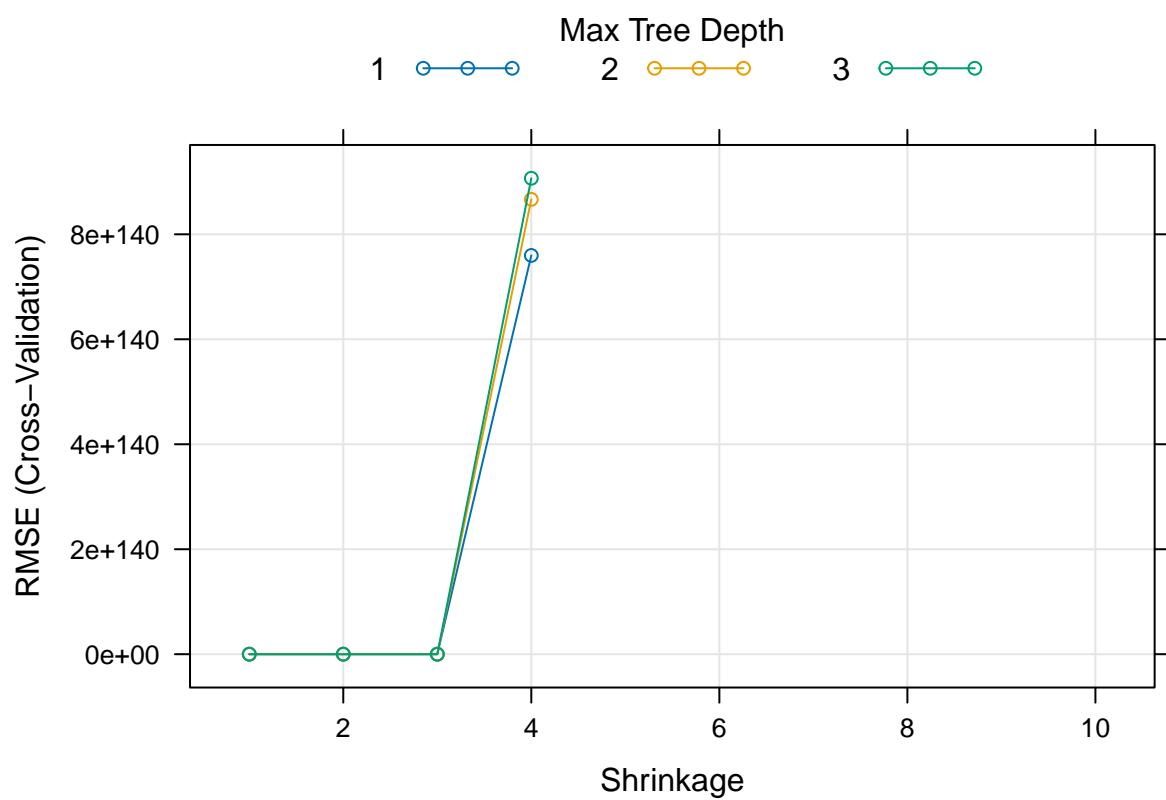
##run models
afa_trans_tune_1 = ozone_tune_models(afa_8_hr_trans, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
                                         lassoLambda_AFA_trans, boostLambda_AFA_trans, annLambda_AFA_trans)

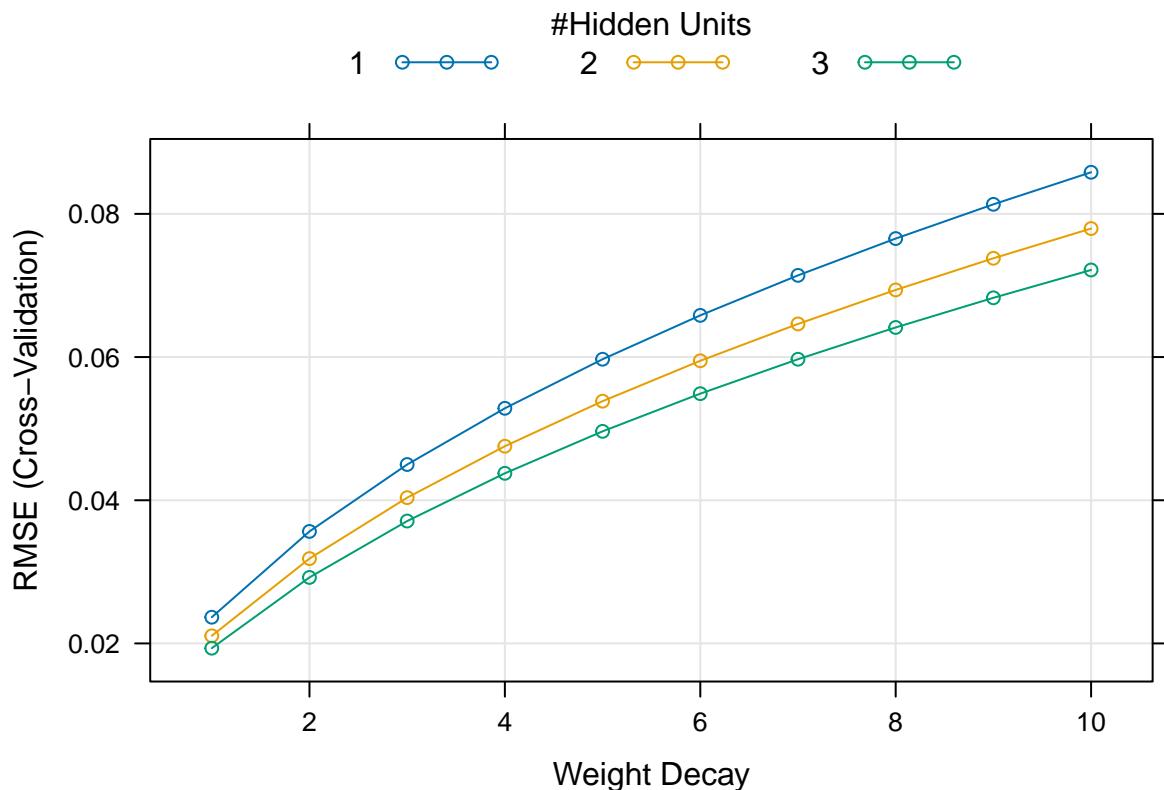
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 4.53999297624849e-05"
## [1] "2023-12-09 23:10:20 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 1"
## [1] "2023-12-09 23:10:44 MST"
## [1] "Artificial Neural Net has been fitted"
```

```
## [1] "Best lambda for Artificial Neural Net is: 1"  
## [1] "2023-12-09 23:10:48 MST"
```







```

## [1] 4.539993e-05 1.000000e+00 1.000000e+00

#AFA AFA_transformed tuning #2

###Updated Values
#LASSO
lassoLambda_AFA_trans = exp(-150:-35/10) #returned to get smaller values
#GBM
boostLambda_AFA_trans = seq(.1,afa_trans_tune_1[2],.1) #best tune is 1. Examine 10x finer scale
#ANN
annLambda_AFA_trans = seq(.1,afa_trans_tune_1[3],.1) #best tune is 1. Examine 10x finer scale

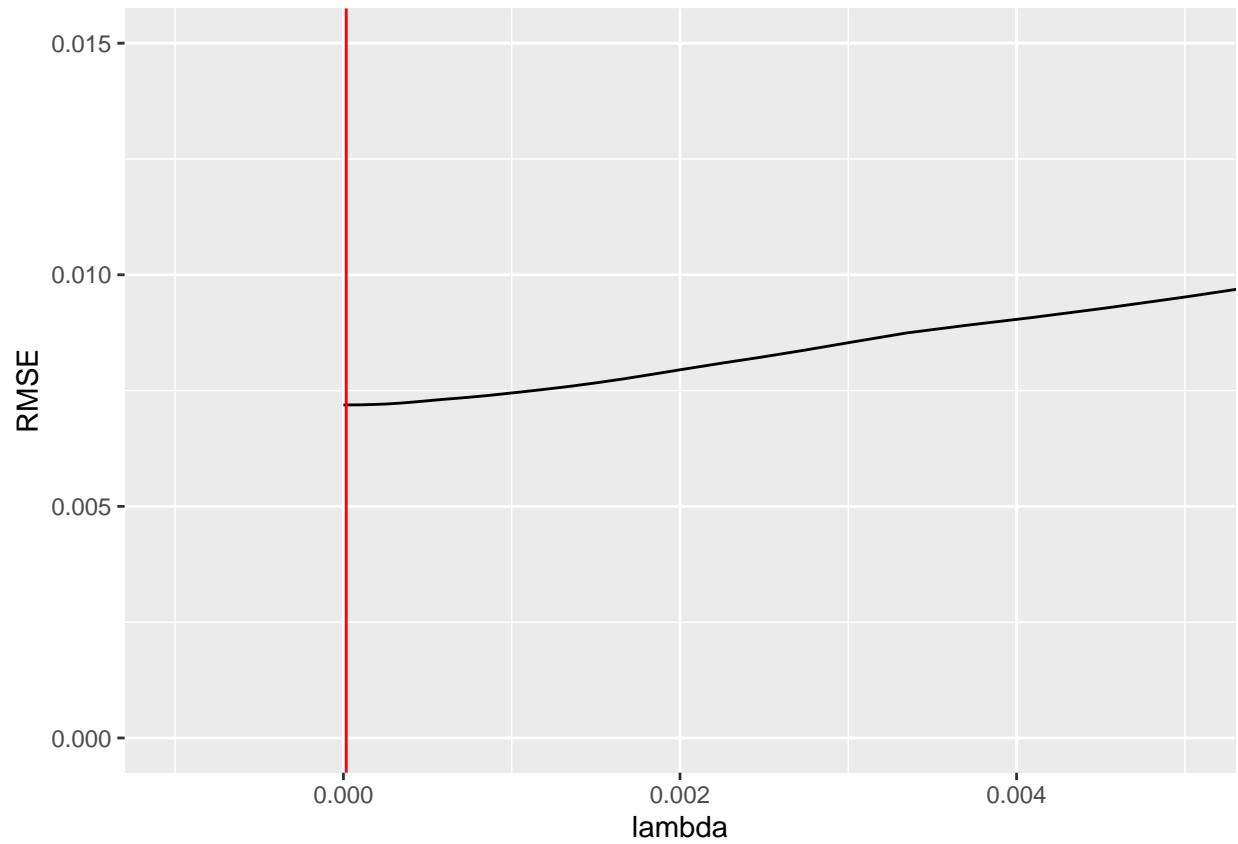
##run models
afa_trans_tune_2 = ozone_tune_models(afa_8_hr_trans, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean")

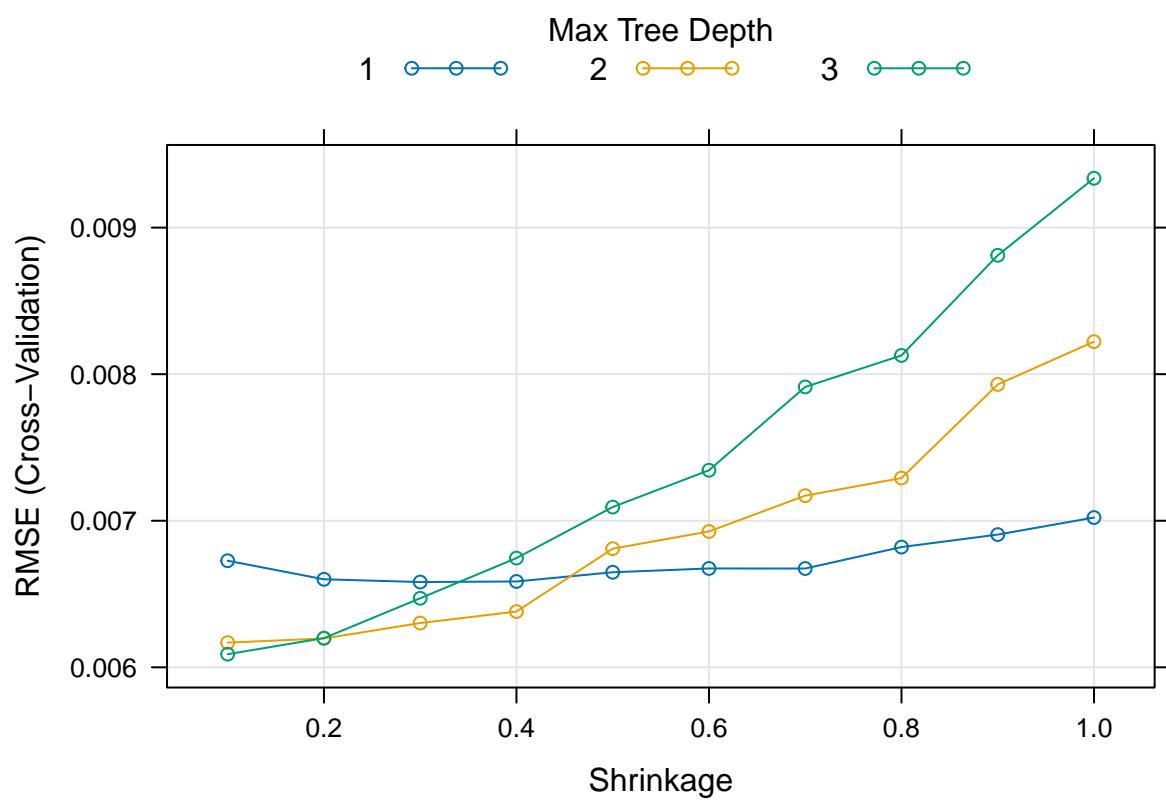
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

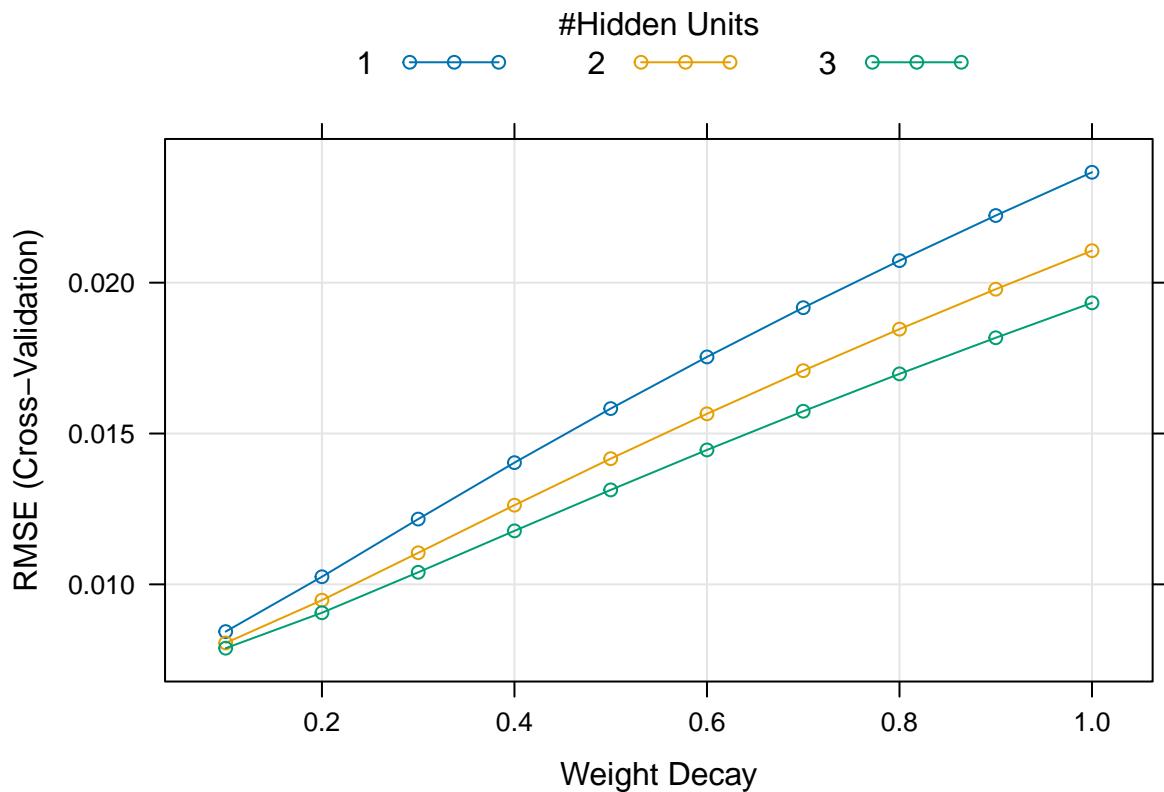
## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.67017007902457e-05"
## [1] "2023-12-09 23:10:49 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.1"
## [1] "2023-12-09 23:11:13 MST"
## [1] "Artificial Neural Net has been fitted"

```

```
## [1] "Best lambda for Artificial Neural Net is: 0.1"  
## [1] "2023-12-09 23:11:16 MST"
```







```

## [1] 1.67017e-05 1.00000e-01 1.00000e-01

#AFA transformed tuning #3

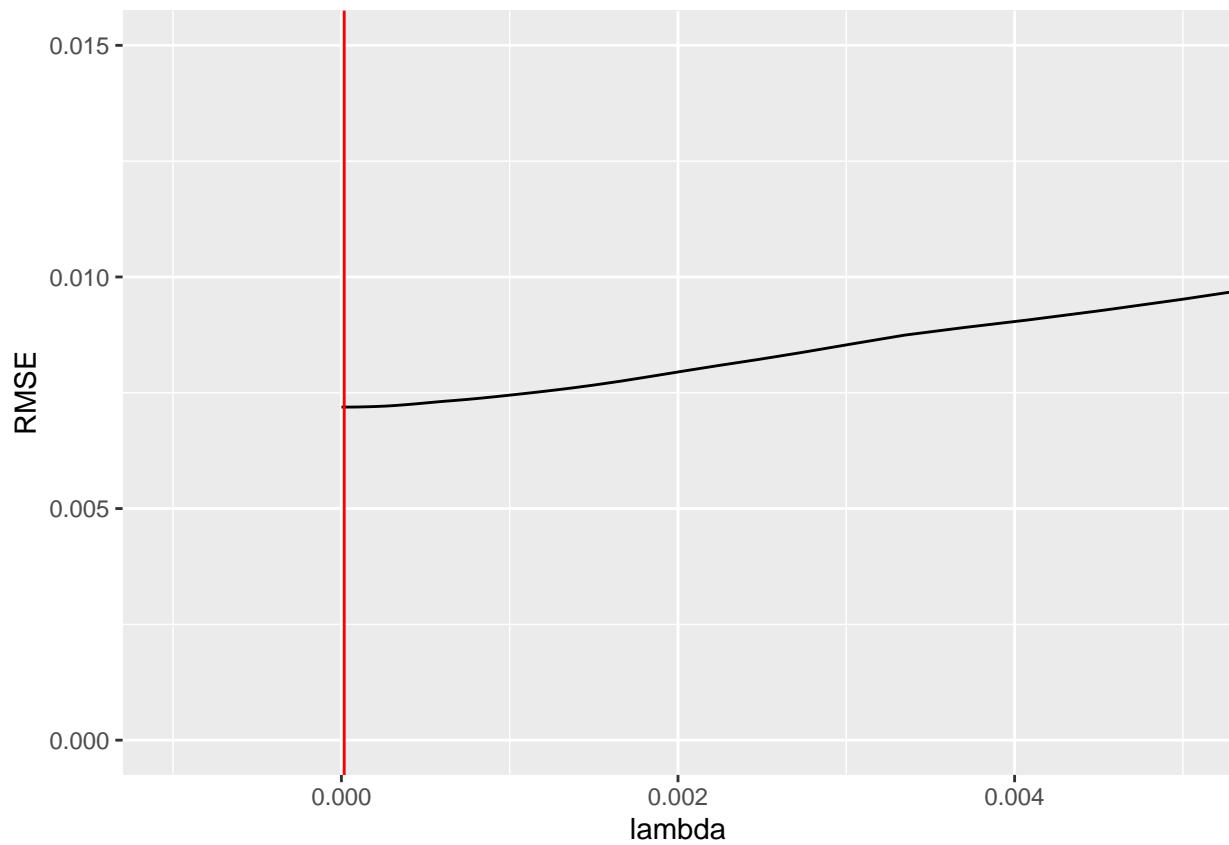
####Updated Values
#Keep LASSO values
boostLambda_AFA_trans = seq(.07,afa_trans_tune_2[2]+.07,.01) #best tune is 0.1, lower & upper bound extended
annLambda_AFA_trans = seq(.01,afa_trans_tune_2[3],.01) #best tune is 0.1, lower bound extended. Examining the results, it appears that the best lambda for the ANN is 0.1, which is the same as the LASSO model.

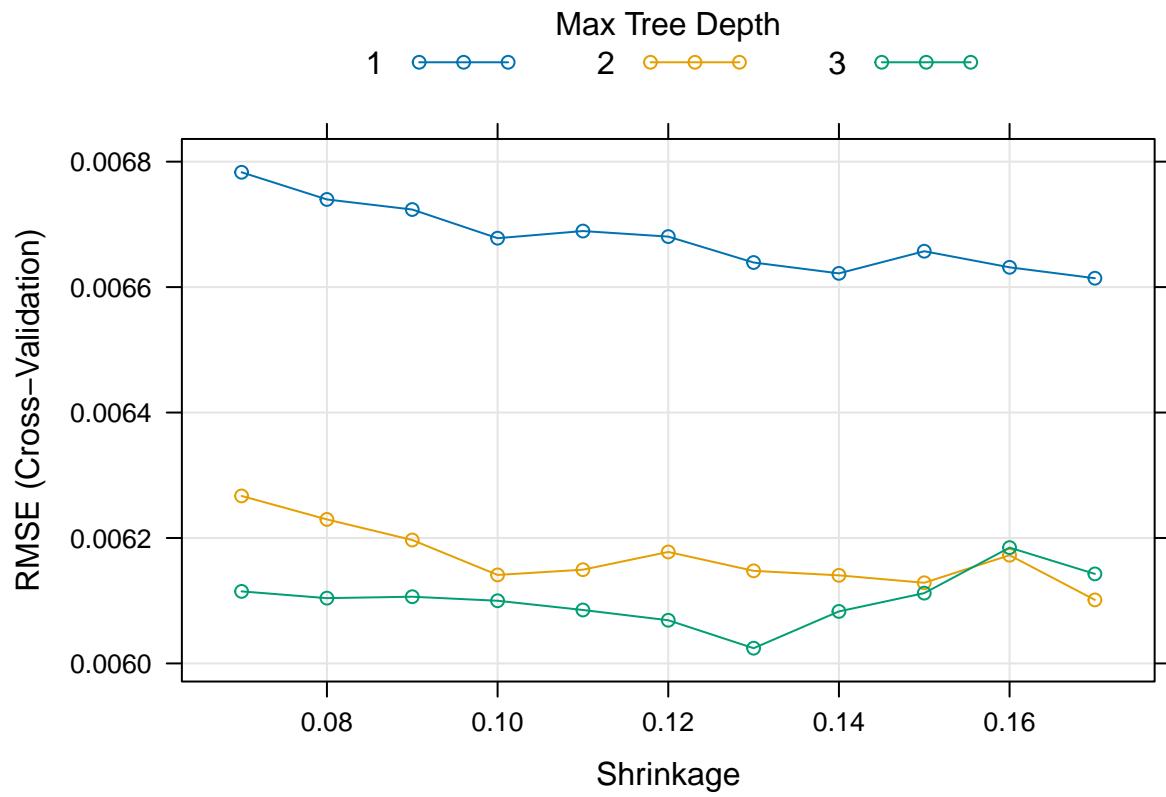
afa_trans_tune_3 = ozone_tune_models(afa_8_hr_trans, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean")

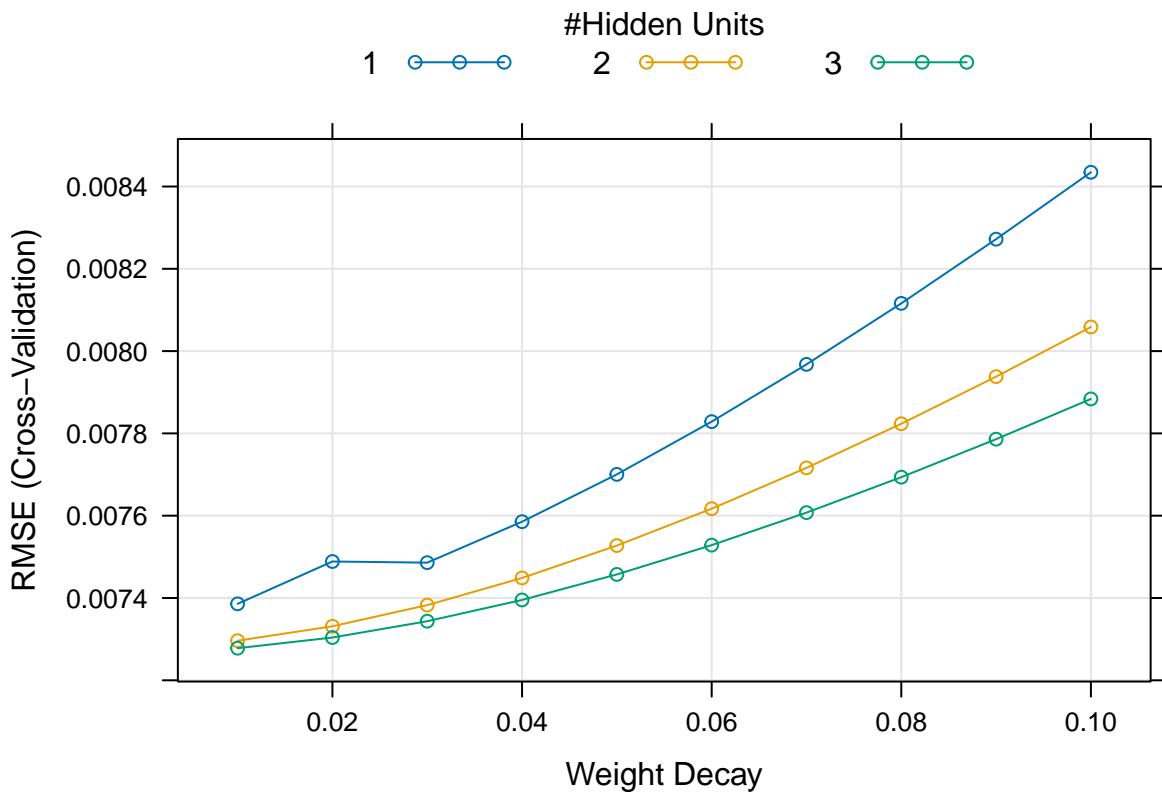
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.67017007902457e-05"
## [1] "2023-12-09 23:11:17 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.13"
## [1] "2023-12-09 23:11:43 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.01"
## [1] "2023-12-09 23:11:47 MST"

```







```

## [1] 1.67017e-05 1.30000e-01 1.00000e-02

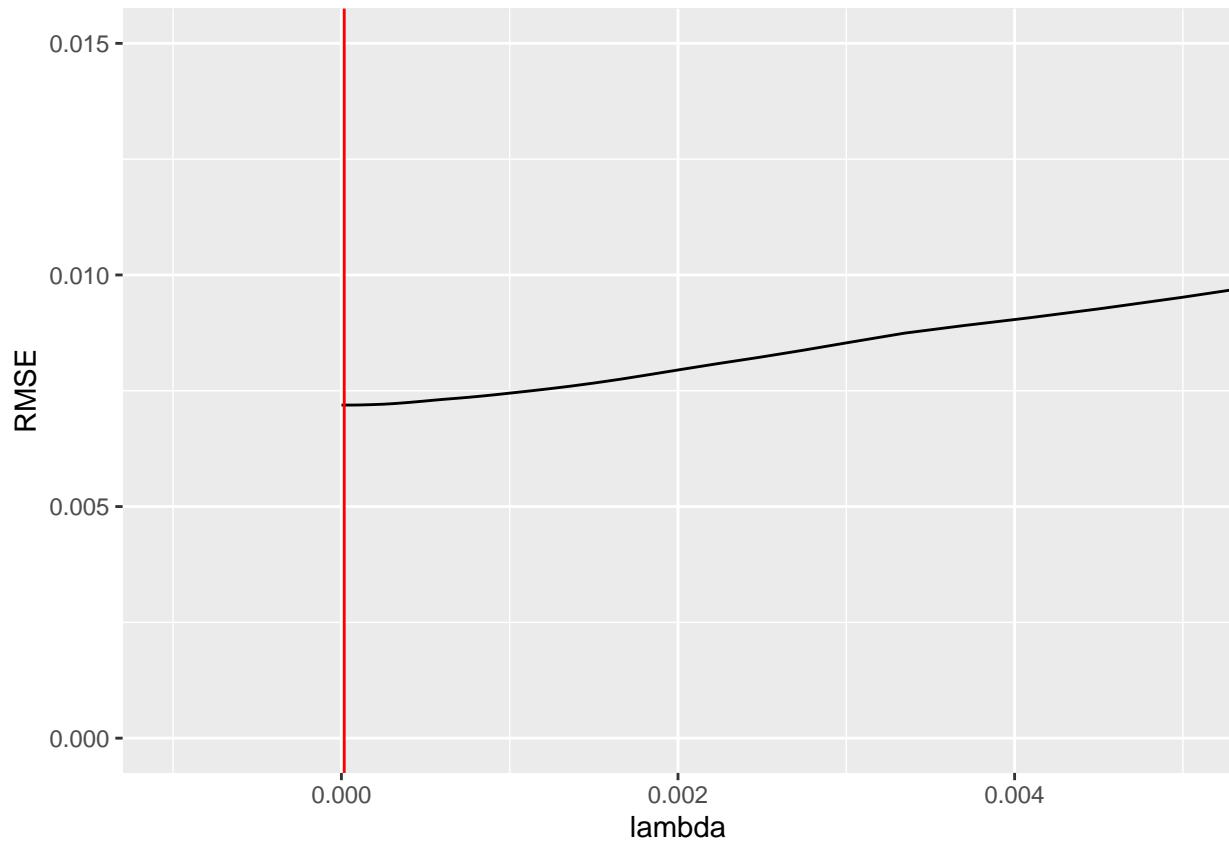
#AFA transformed tuning #4 (final Tune)

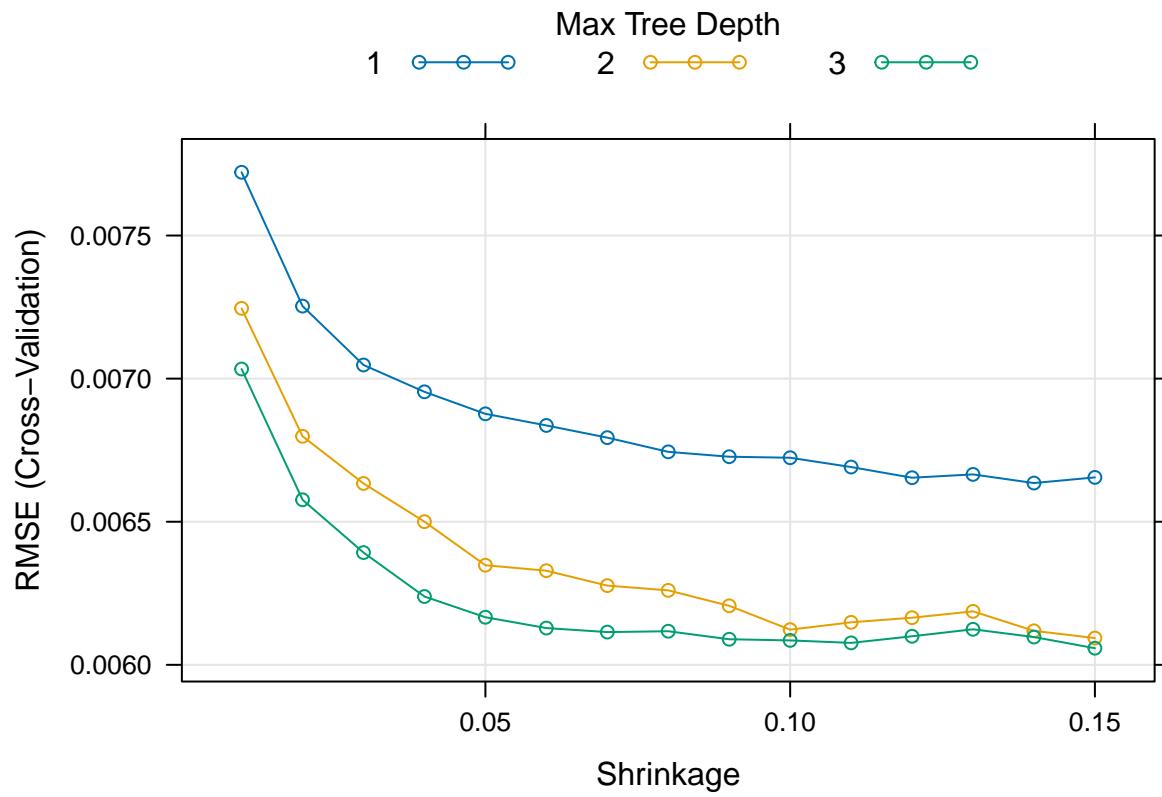
### Updated Values
boostLambda_AFA_trans = seq(.01,afa_trans_tune_2[2]+.05,.01) #shorten for final model
annLambda_AFA_trans = seq(.001,afa_trans_tune_3[3],.001) #best tune is 0.01, lower bound examined. Example
afa_trans_tune_4 = ozone_tune_models(afa_8_hr_trans, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean")

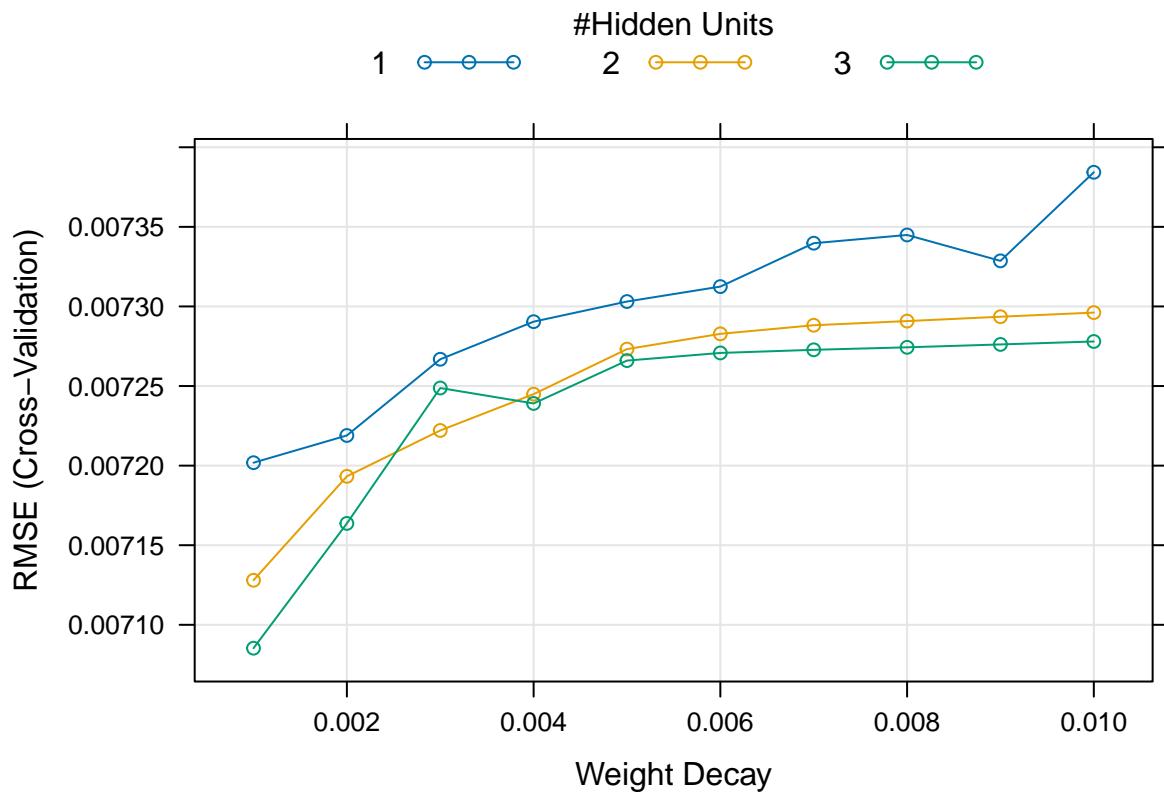
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.67017007902457e-05"
## [1] "2023-12-09 23:11:48 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.15"
## [1] "2023-12-09 23:12:24 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.001"
## [1] "2023-12-09 23:12:32 MST"

```





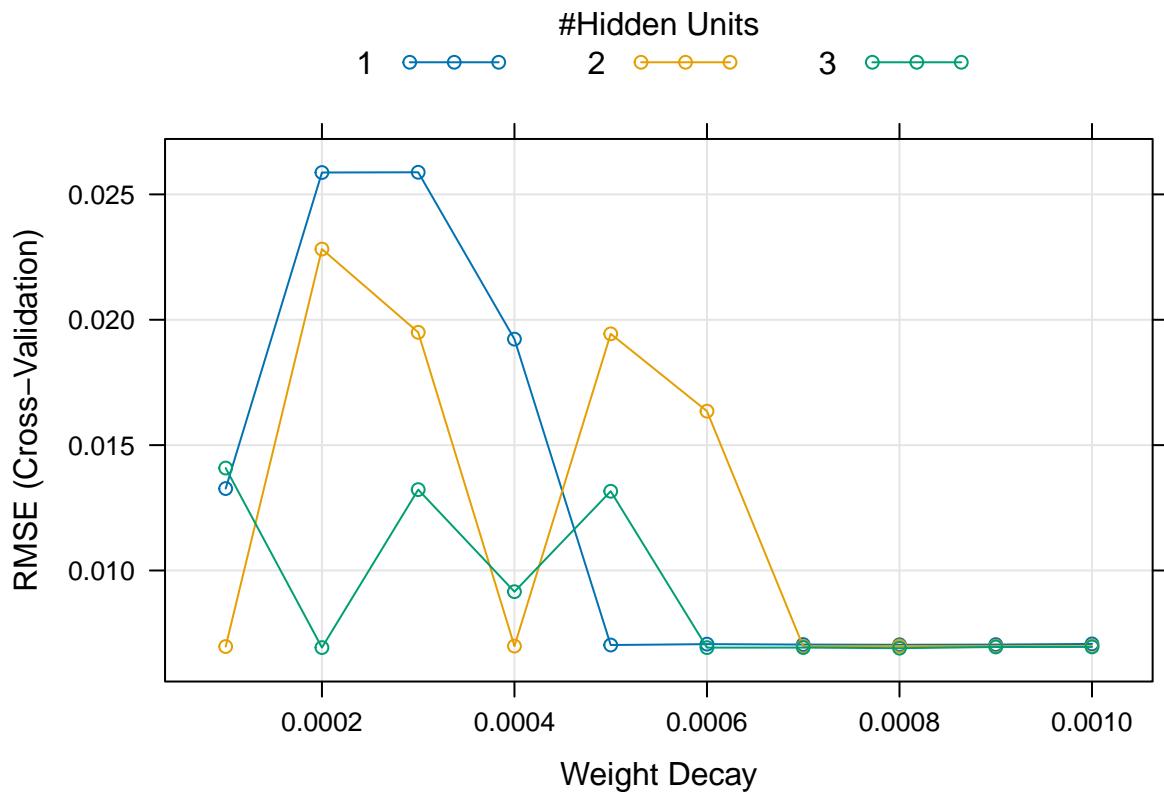


```
## [1] 1.67017e-05 1.50000e-01 1.00000e-03
```

```
#final ann finnessing
annLambda_AFA_trans = seq(.0002,.002,.0002) #best tune is continuously smaller, lower bound extended.
ann = ozone_tuning_ANN(afa_8_hr, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",annLambda_AFA)

## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 8e-04"
## [1] "2023-12-09 23:12:41 MST"
```

```
plot(ann)
```



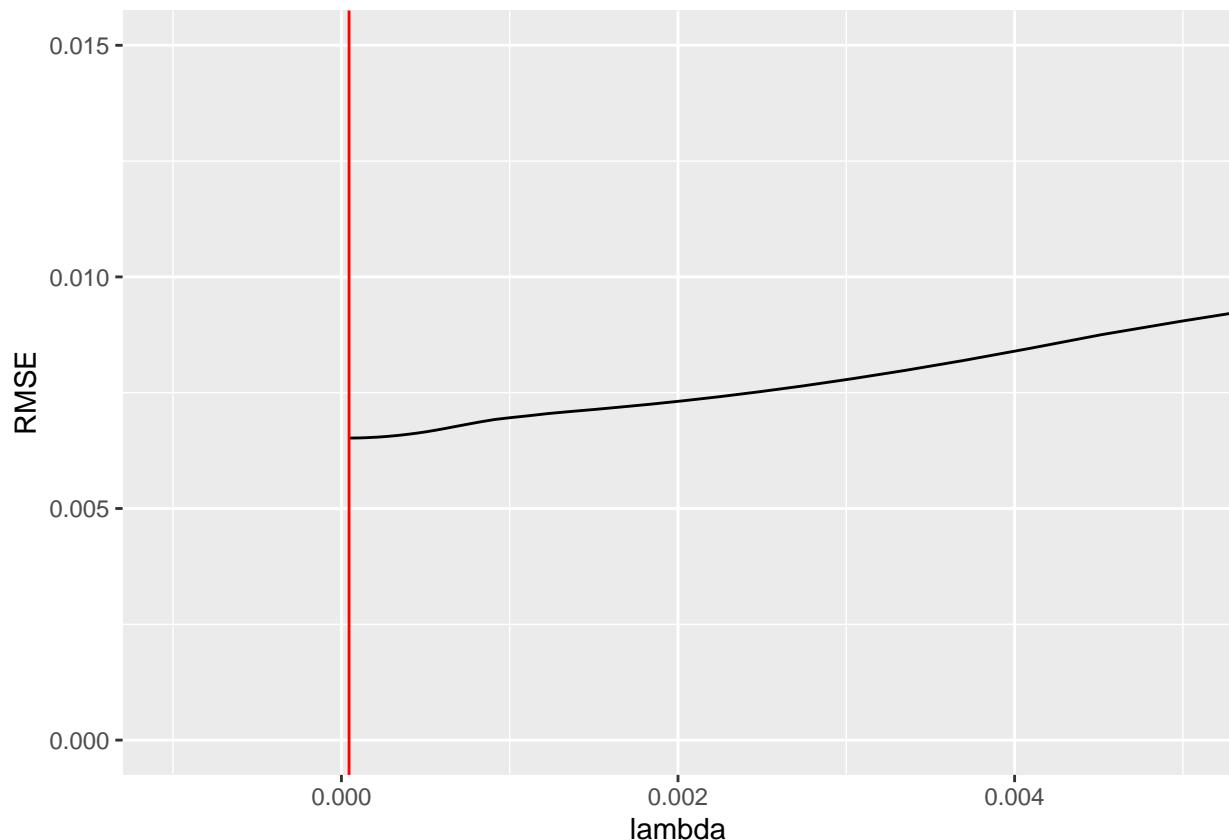
```
#MAN tuning #1
```

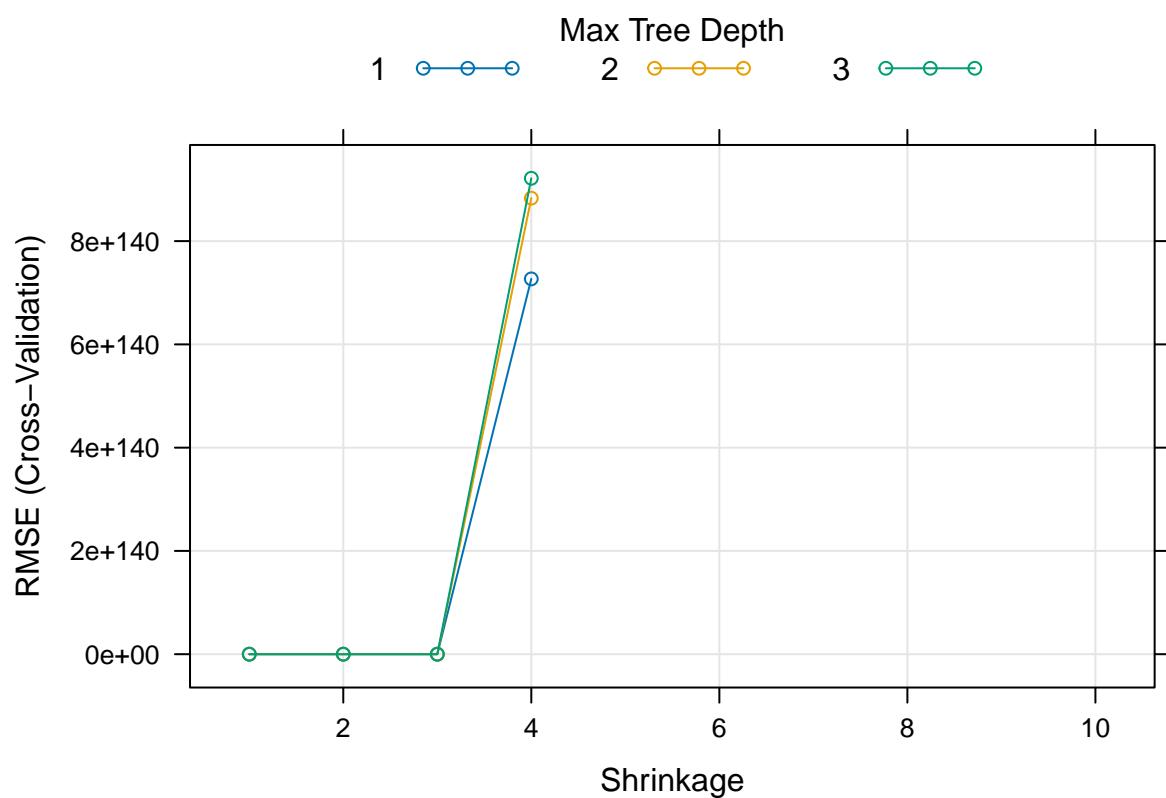
```
#dataset
man_8_hr = ozone_8_hr[-c(1)] #use only MAN measurements
####Starting Values
#LASSO
lassoLambda_MAN = exp(-100:15/10) #initial search space
#GBM
boostLambda_MAN = seq(1,10,1) #examine 1-10 for initial search
#ANN
annLambda_MAN = seq(1,10,1) #examine 1-10 for initial search
##run models
man_tune_1 = ozone_tune_models(man_8_hr, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",
                                lassoLambda_MAN, boostLambda_MAN, annLambda_MAN)

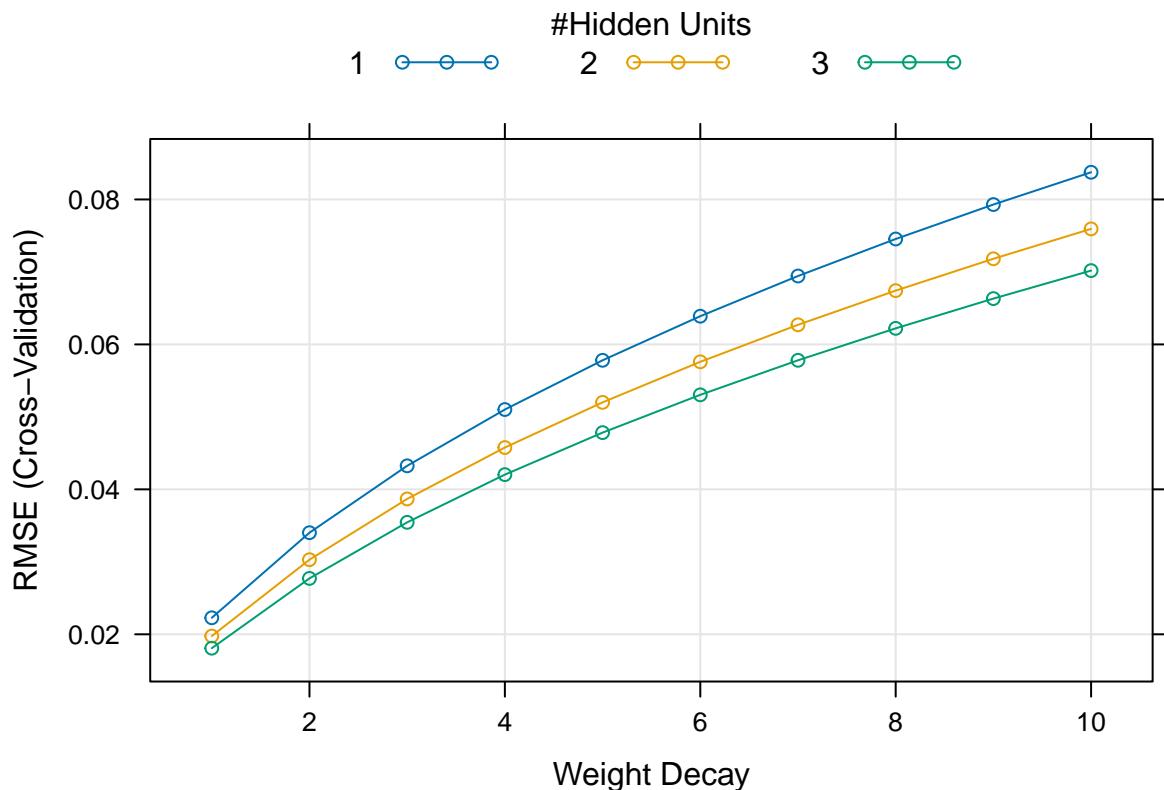
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 4.53999297624849e-05"
## [1] "2023-12-09 23:12:41 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 1"
## [1] "2023-12-09 23:13:14 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 1"
```

```
## [1] "2023-12-09 23:13:19 MST"
```







```

## [1] 4.539993e-05 1.000000e+00 1.000000e+00

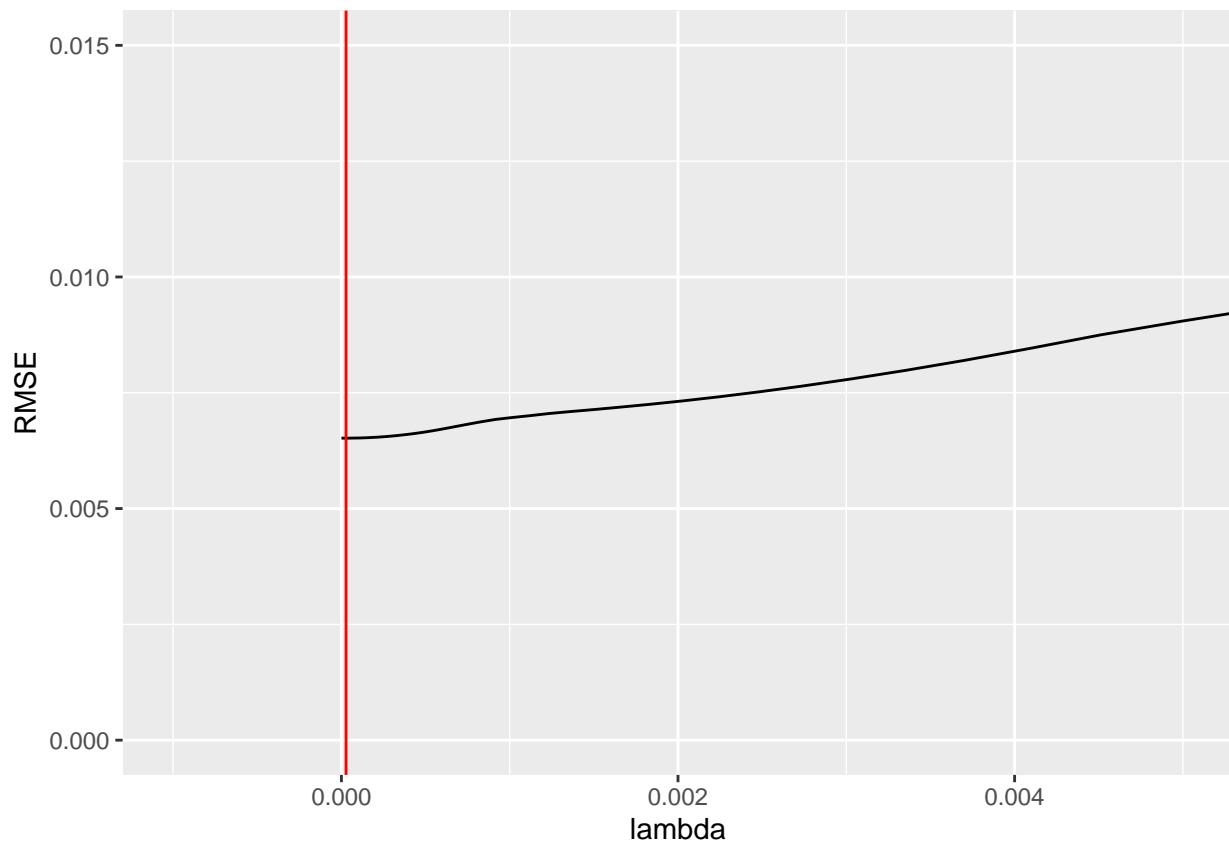
#MAN tuning #2

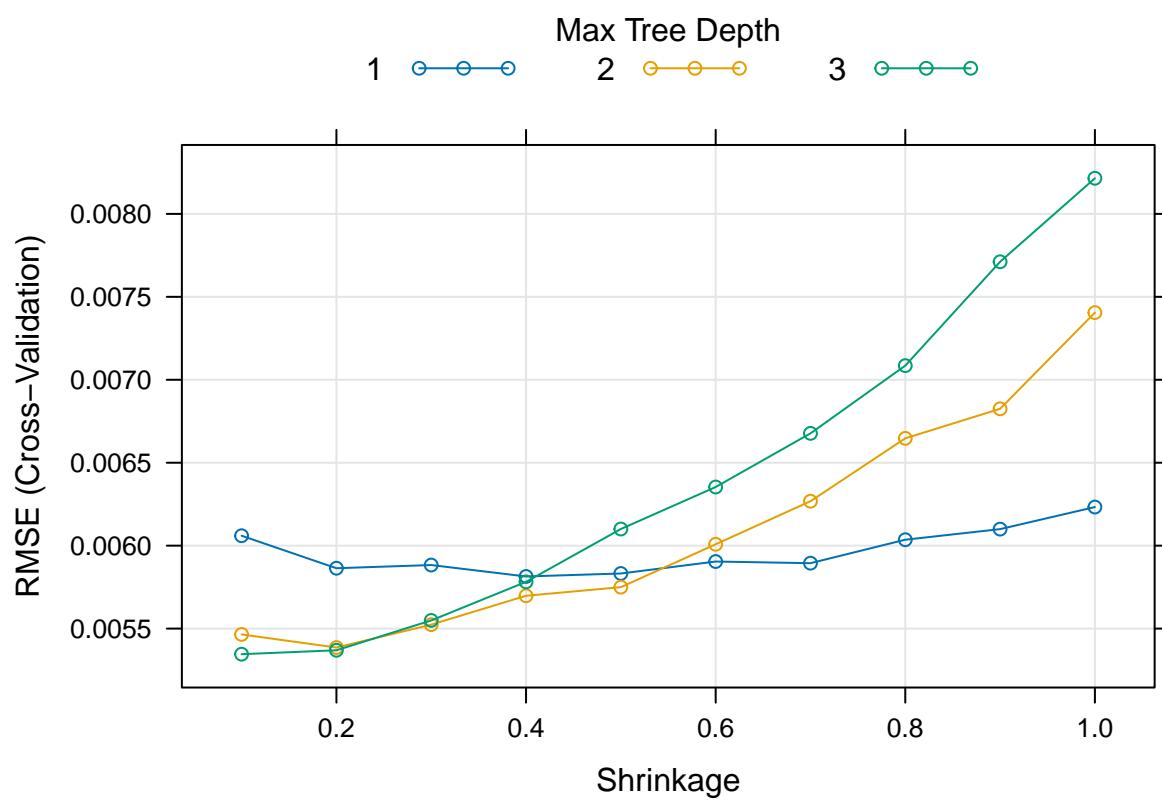
### Updated Values
lassoLambda_MAN = exp(-150:-35/10) #returned to get smaller values
boostLambda_MAN = seq(.1,man_tune_1[2],.1) #best tune is 1, lower bound. Examine 10x finer scale
annLambda_MAN = seq(.1,man_tune_1[3],.1) #best tune is 1, lower bound. Examine 10x finer scale
##run models
man_tune_2 = ozone_tune_models(man_8_hr, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean", lassoLambda_MAN)

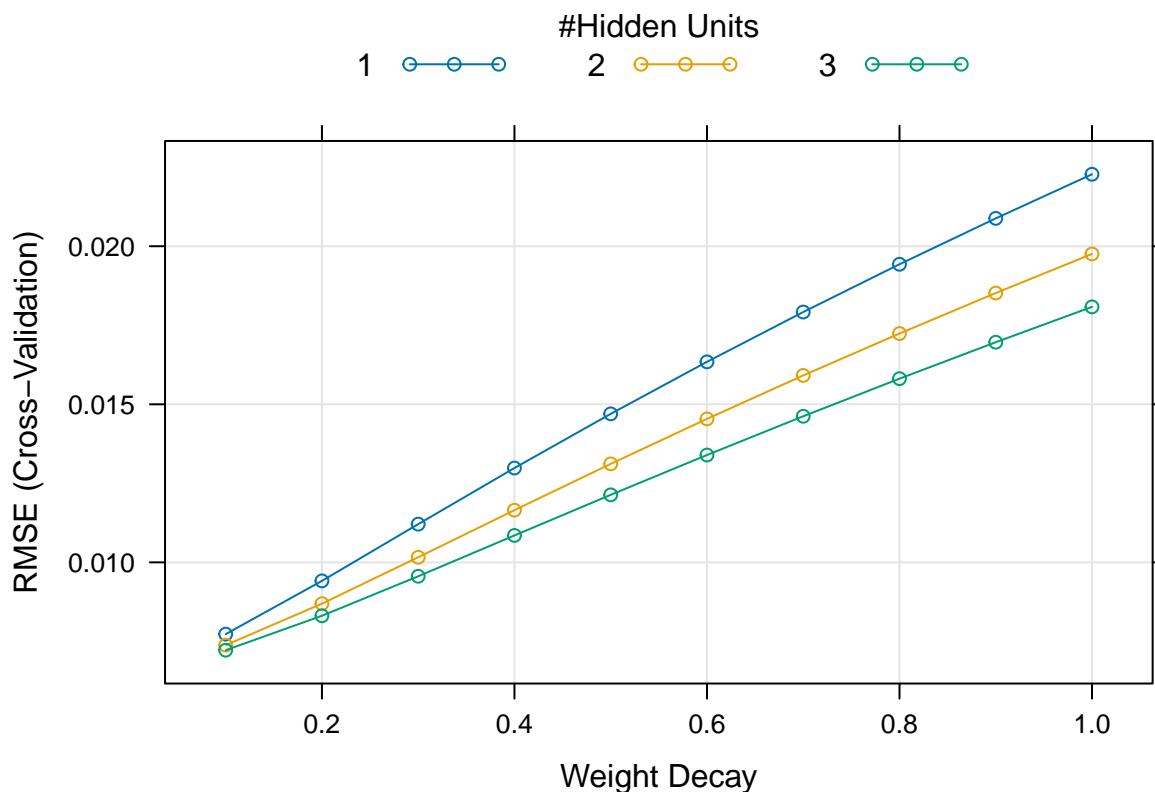
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 2.75364493497472e-05"
## [1] "2023-12-09 23:13:19 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.1"
## [1] "2023-12-09 23:13:52 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.1"
## [1] "2023-12-09 23:13:56 MST"

```







```

## [1] 2.753645e-05 1.000000e-01 1.000000e-01

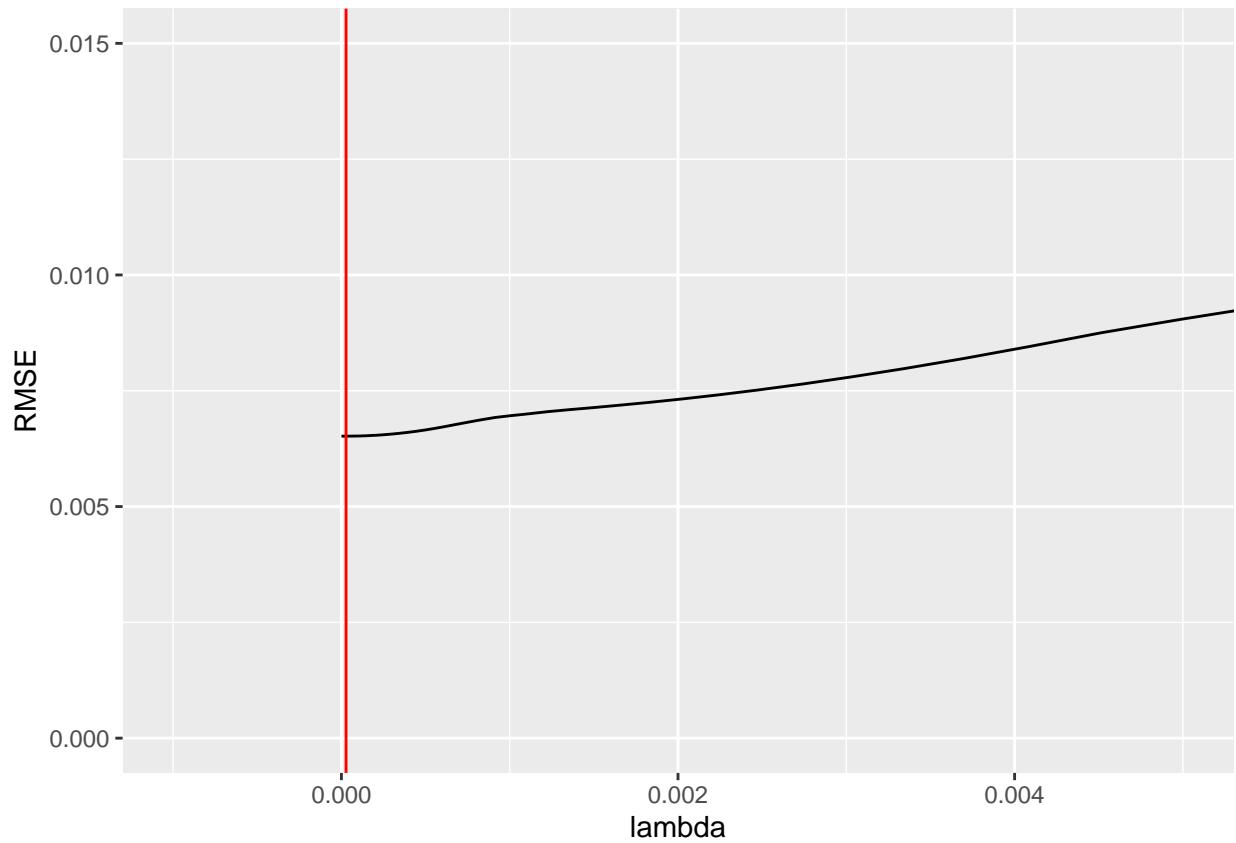
#MAN tuning #3

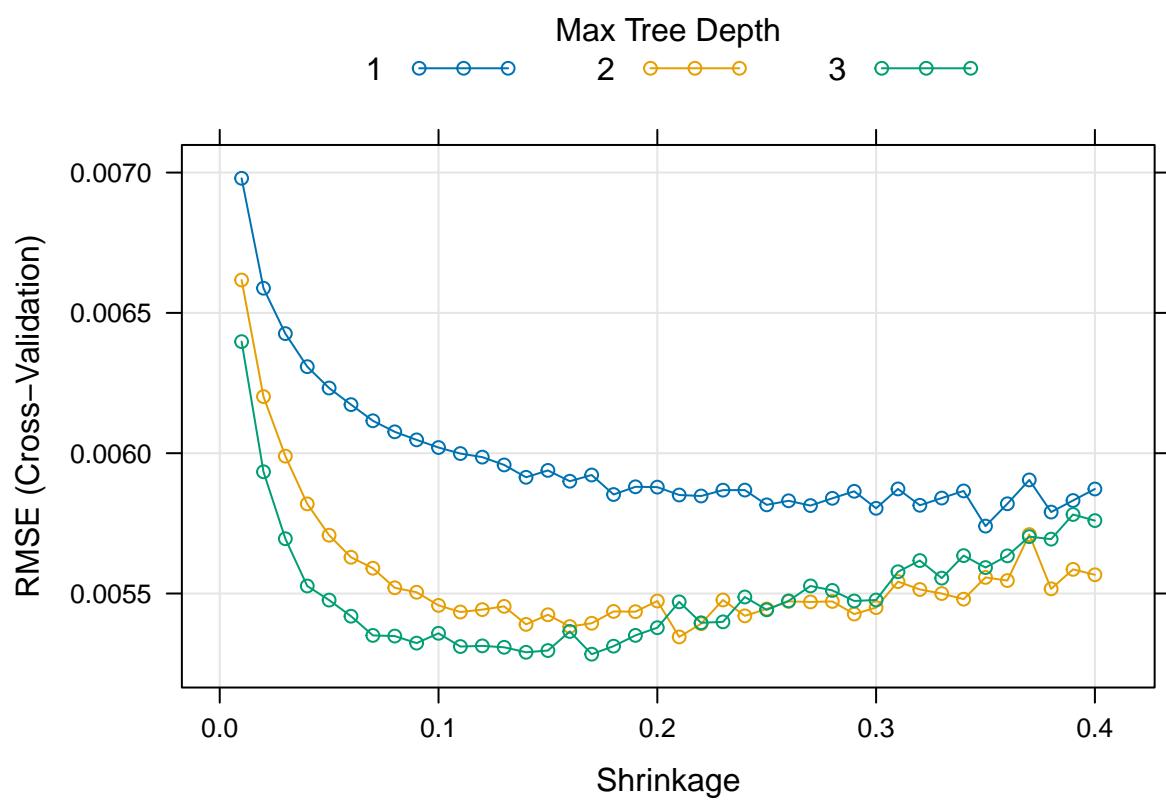
####Updated Values -
#Keep LASSO lambda since best tune lies within distribution of values
boostLambda_MAN = seq(.01,man_tune_2[2]+.3,.01) #best tune is 0.1, lower & upper bound extended. Examine 10x finer
annLambda_MAN = seq(.01,man_tune_2[3],.01) #best tune is 0.1, lower bound extended. Examine 10x finer
#run
man_tune_3 = ozone_tune_models(man_8_hr, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean", lassoLambda

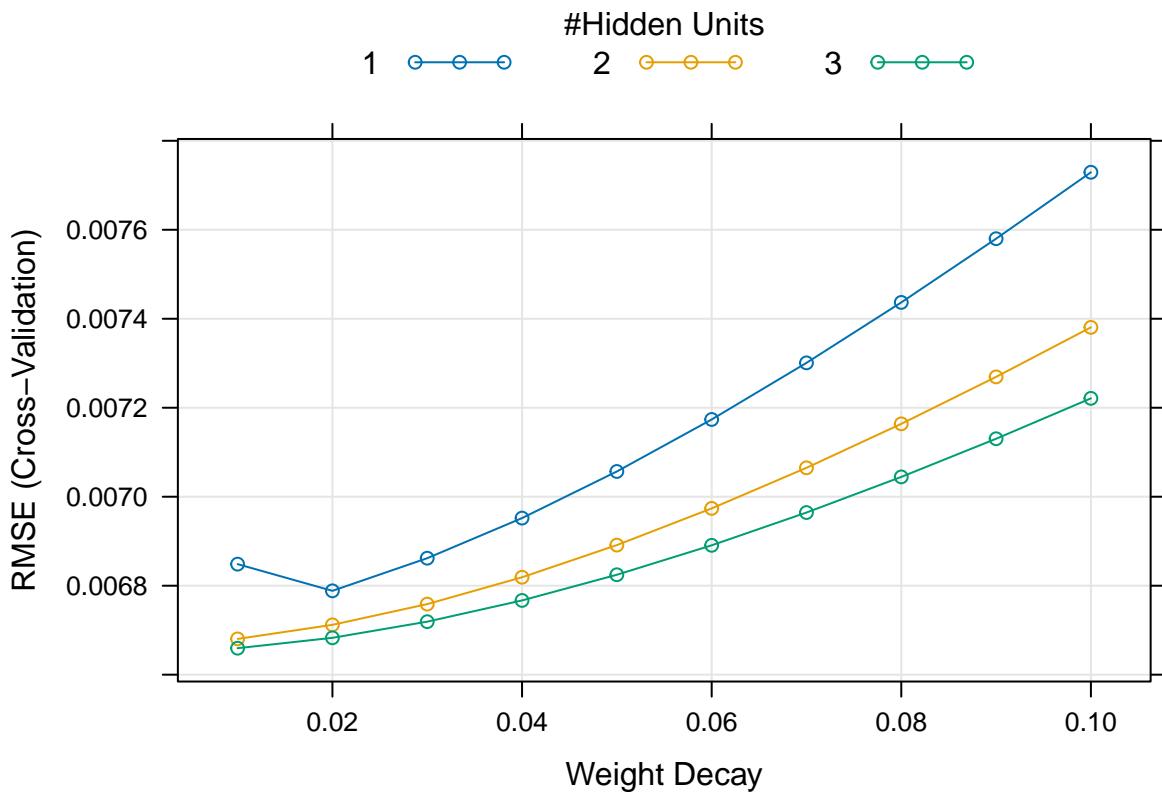
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 2.75364493497472e-05"
## [1] "2023-12-09 23:13:57 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.17"
## [1] "2023-12-09 23:15:55 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.01"
## [1] "2023-12-09 23:16:00 MST"

```







```

## [1] 2.753645e-05 1.700000e-01 1.000000e-02

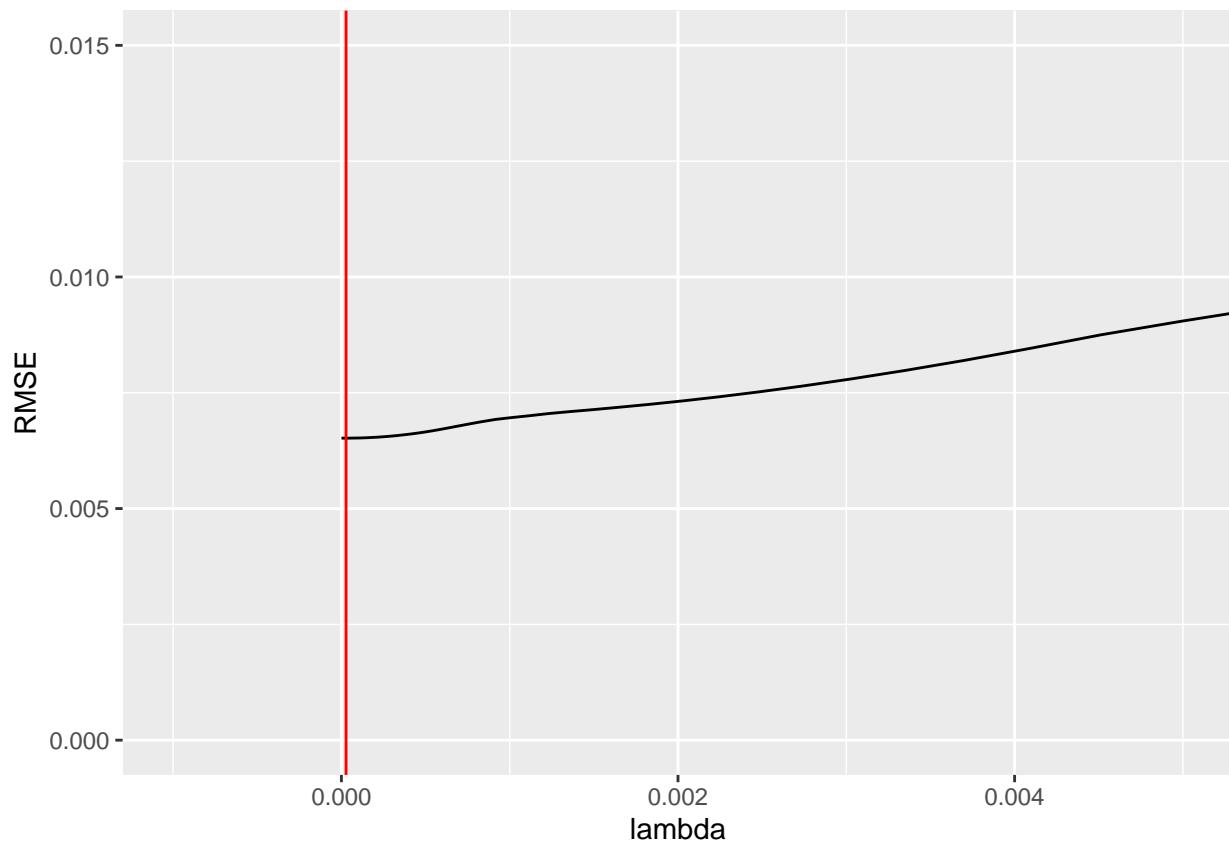
#MAN tuning #4 (final Tune)

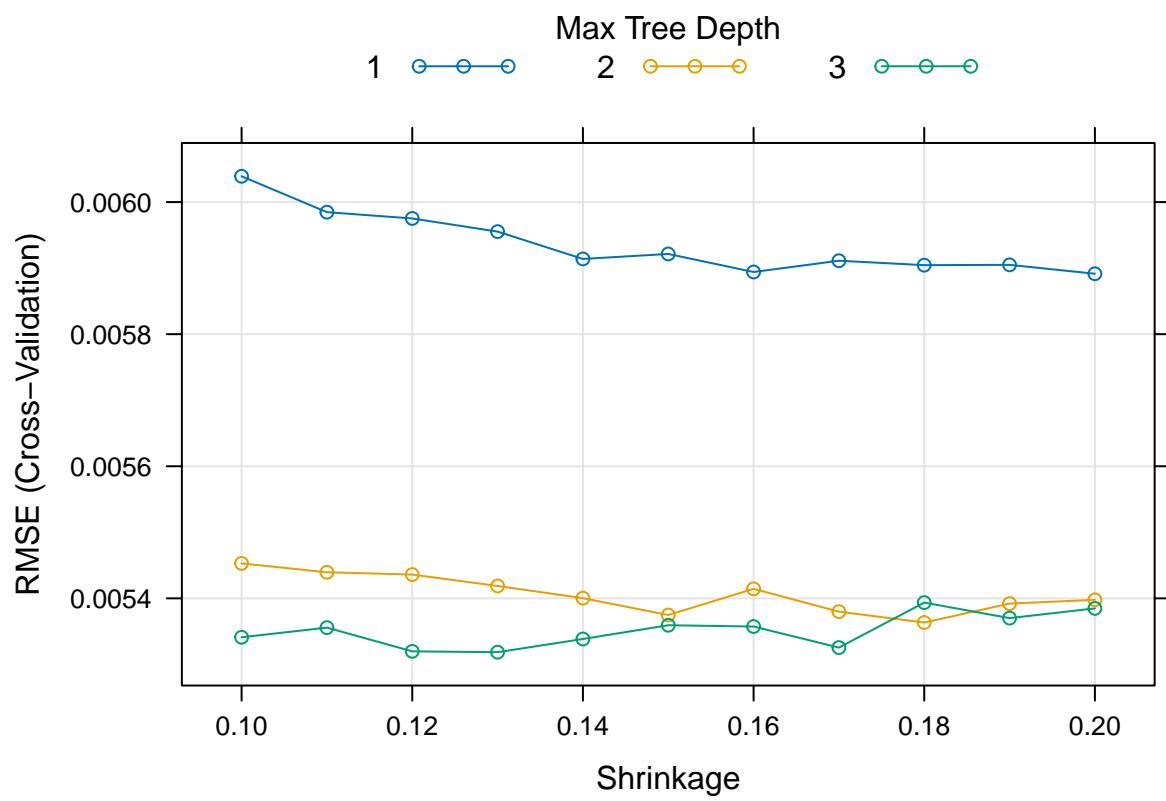
####Updated Values
boostLambda_MAN = seq(.1,.2,.01) #shorten scale for final models
annLambda_MAN = seq(.001,.01,.001) #best tune is continuously smaller, lower bound extended. Examine 1

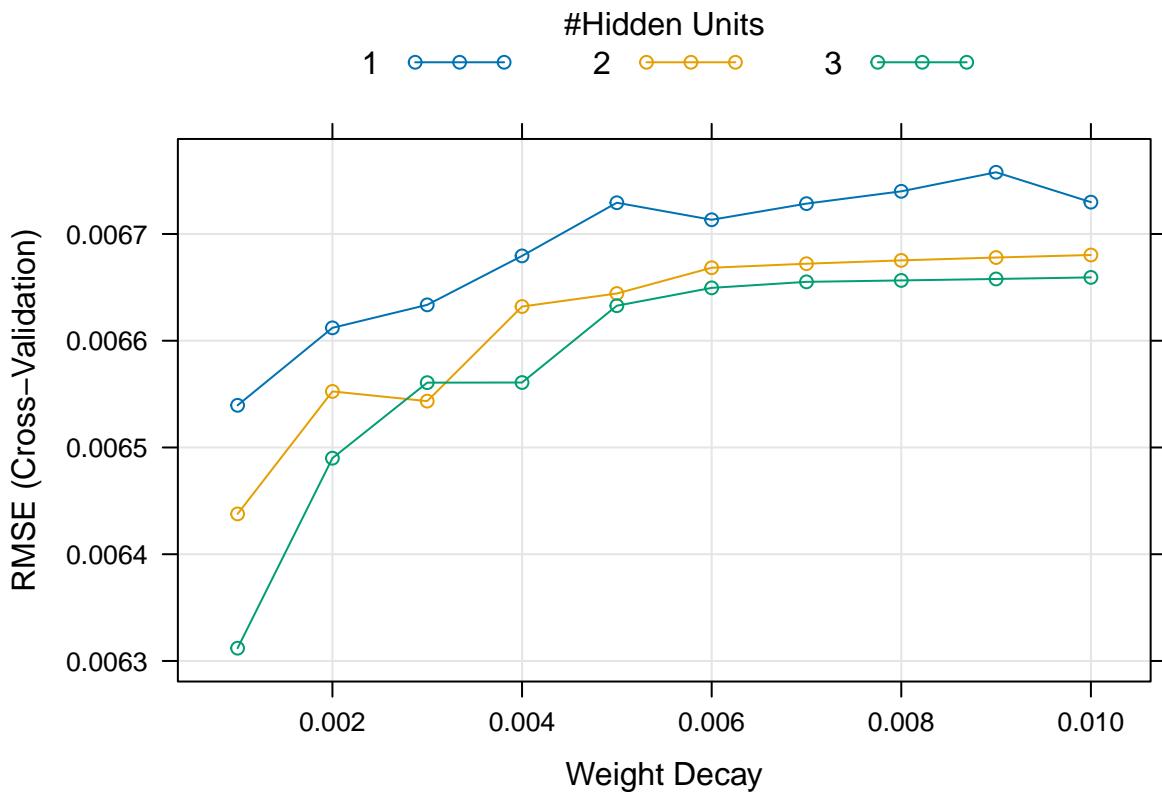
man_tune_4 = ozone_tune_models(man_8_hr, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean", lassoLambda
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 2.75364493497472e-05"
## [1] "2023-12-09 23:16:01 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.13"
## [1] "2023-12-09 23:16:32 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.001"
## [1] "2023-12-09 23:16:41 MST"

```





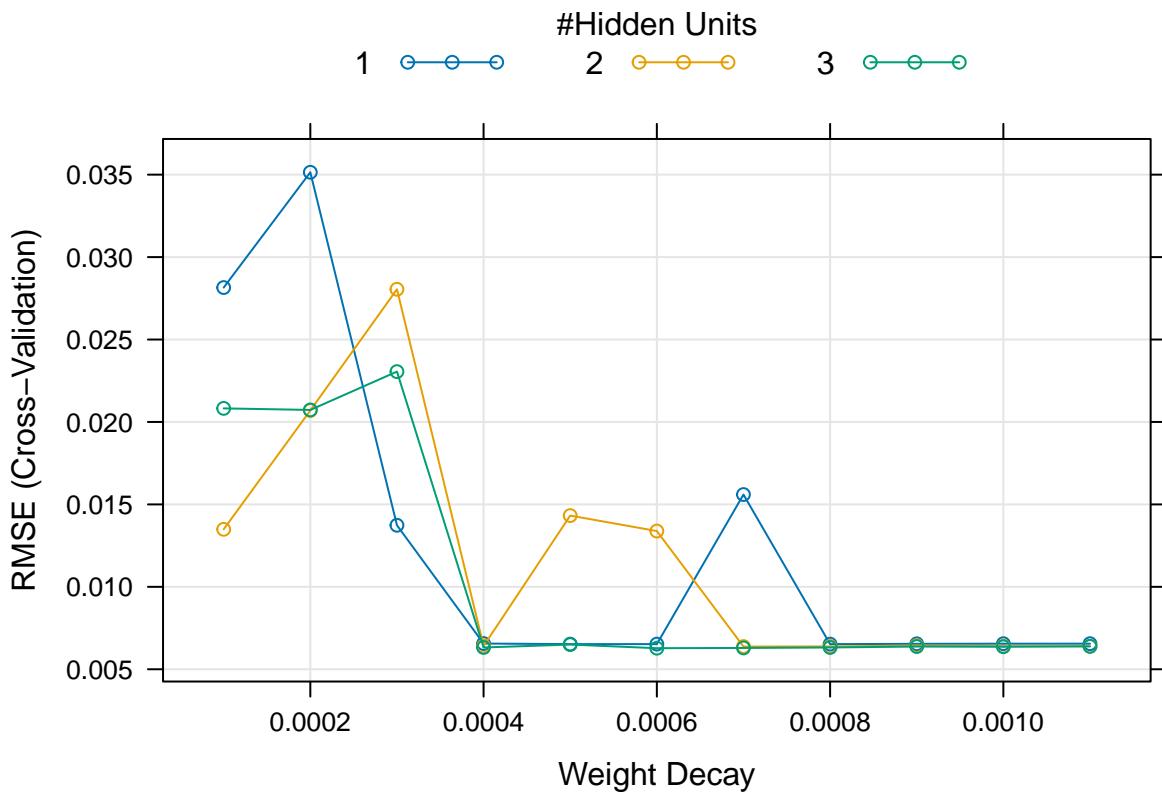


```
## [1] 2.753645e-05 1.300000e-01 1.000000e-03
```

```
#final ann finnessing
annLambda_MAN = seq(.0001,.0011,.0001) #best tune is continuously smaller, lower bound extended. Examining
ann = ozone_tuning_ANN(man_8_hr, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",annLambda_MAN)

## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 6e-04"
## [1] "2023-12-09 23:16:50 MST"
```

```
plot(ann)
```



```
#MAN_transformed tuning #1
```

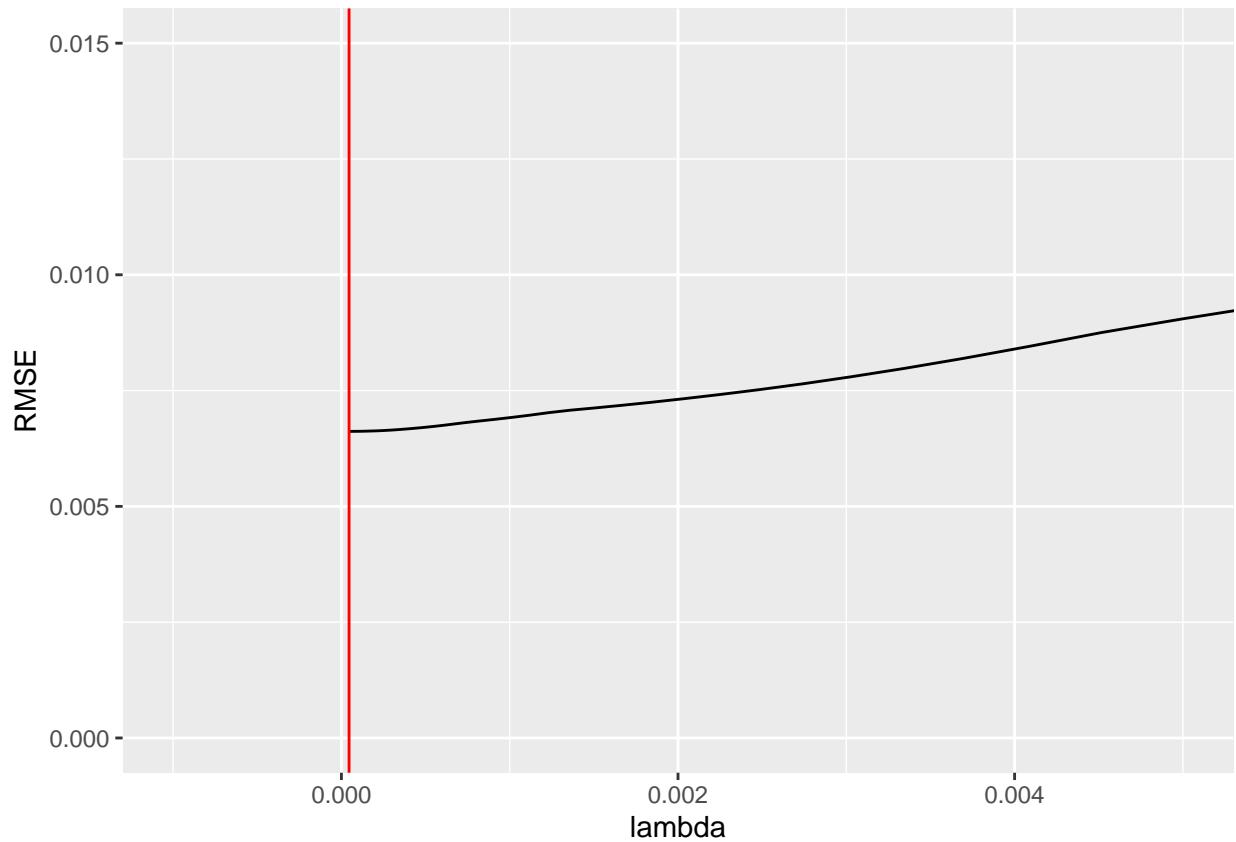
```
#dataset
man_8_hr_trans = ozone_transformed[-c(1)] #use only MAN measurements
####Starting Values
#LASSO
lassoLambda_MAN_trans = exp(-100:15/10) #initial search space
#GBM
boostLambda_MAN_trans = seq(1,10,1) #examine 1-10 for initial search
#ANN
annLambda_MAN_trans = seq(1,10,1) #examine 1-10 for initial search

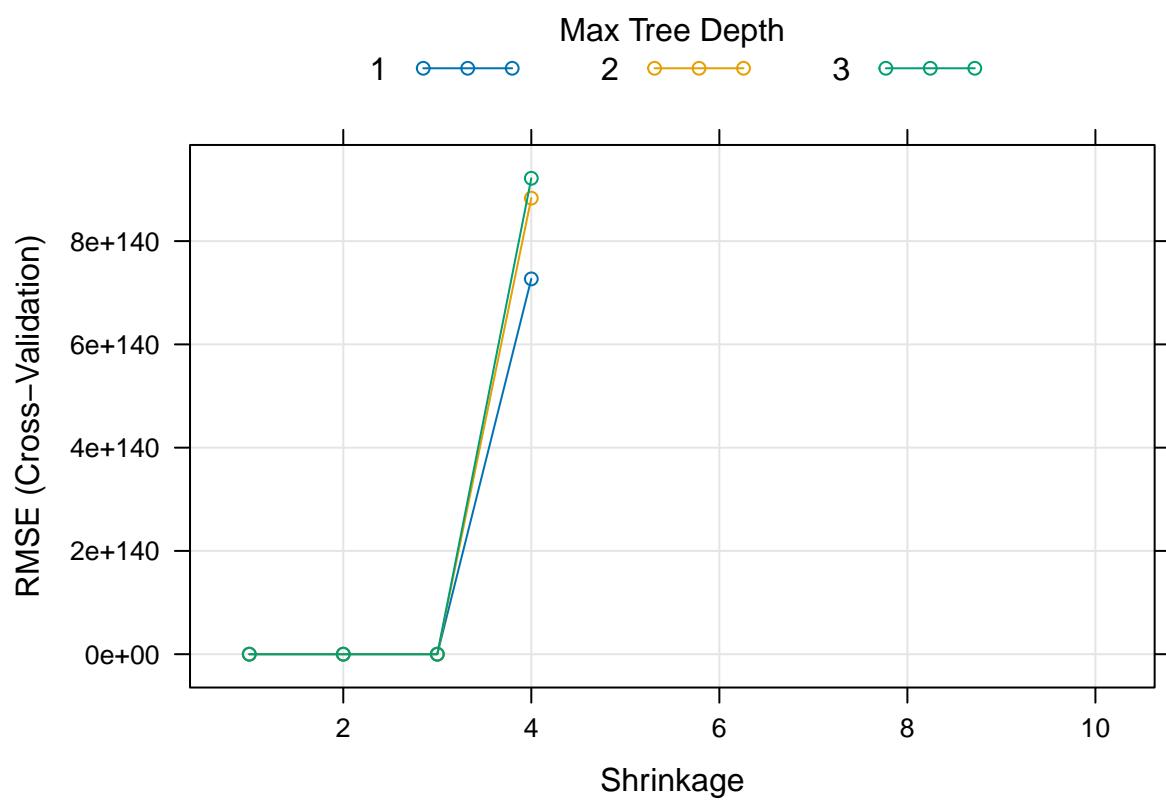
##run models
man_trans_tune_1 = ozone_tune_models(man_8_hr_trans, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
                                         lassoLambda_MAN_trans, boostLambda_MAN_trans, annLambda_MAN_trans)

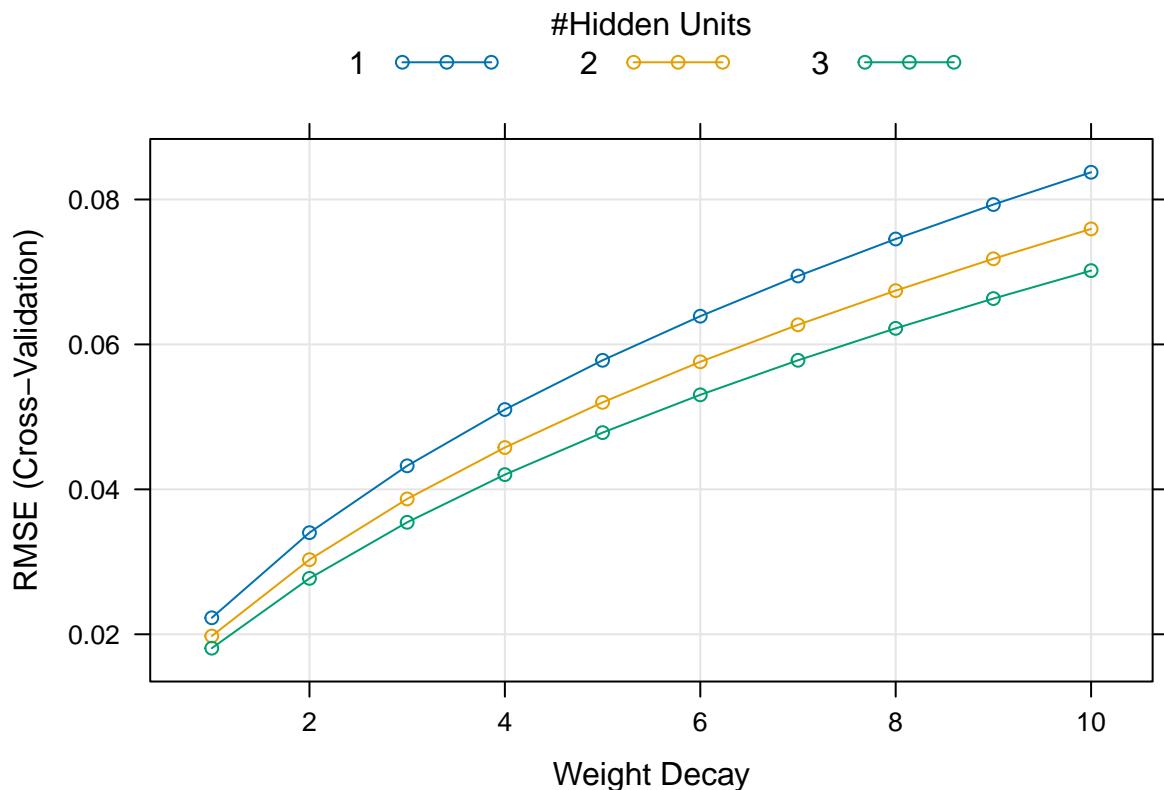
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 4.53999297624849e-05"
## [1] "2023-12-09 23:16:51 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 1"
## [1] "2023-12-09 23:17:15 MST"
## [1] "Artificial Neural Net has been fitted"
```

```
## [1] "Best lambda for Artificial Neural Net is: 1"  
## [1] "2023-12-09 23:17:19 MST"
```







```

## [1] 4.539993e-05 1.000000e+00 1.000000e+00

#MAN transformed tuning #2

###Updated Values
#LASSO
lassoLambda_MAN_trans = exp(-150:-35/10) #returned to get smaller values
#GBM
boostLambda_MAN_trans = seq(.1,man_trans_tune_1[2],.1) #best tune is 1. Examine 10x finer scale
#ANN
annLambda_MAN_trans = seq(.1,man_trans_tune_1[3],.1) #best tune is 1. Examine 10x finer scale

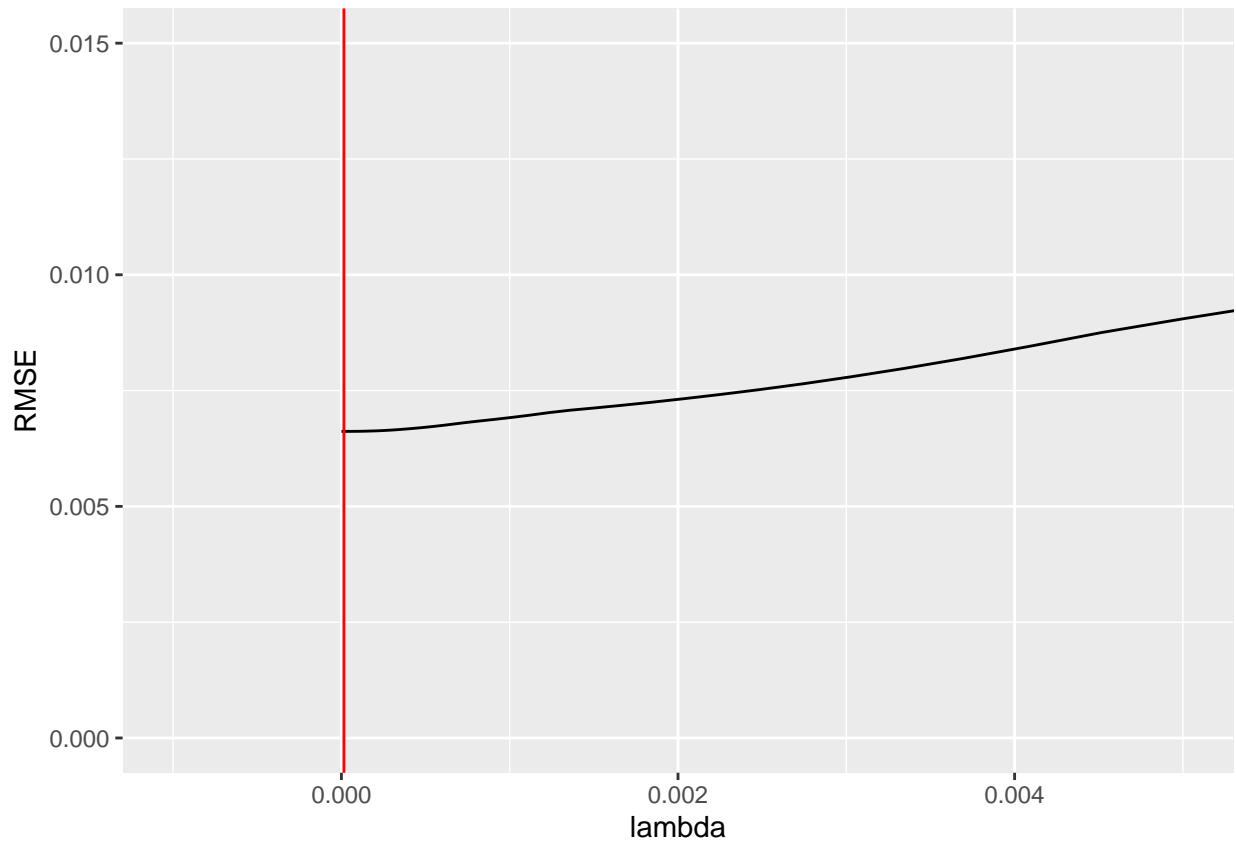
##run models
man_trans_tune_2 = ozone_tune_models(man_8_hr_trans, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean")

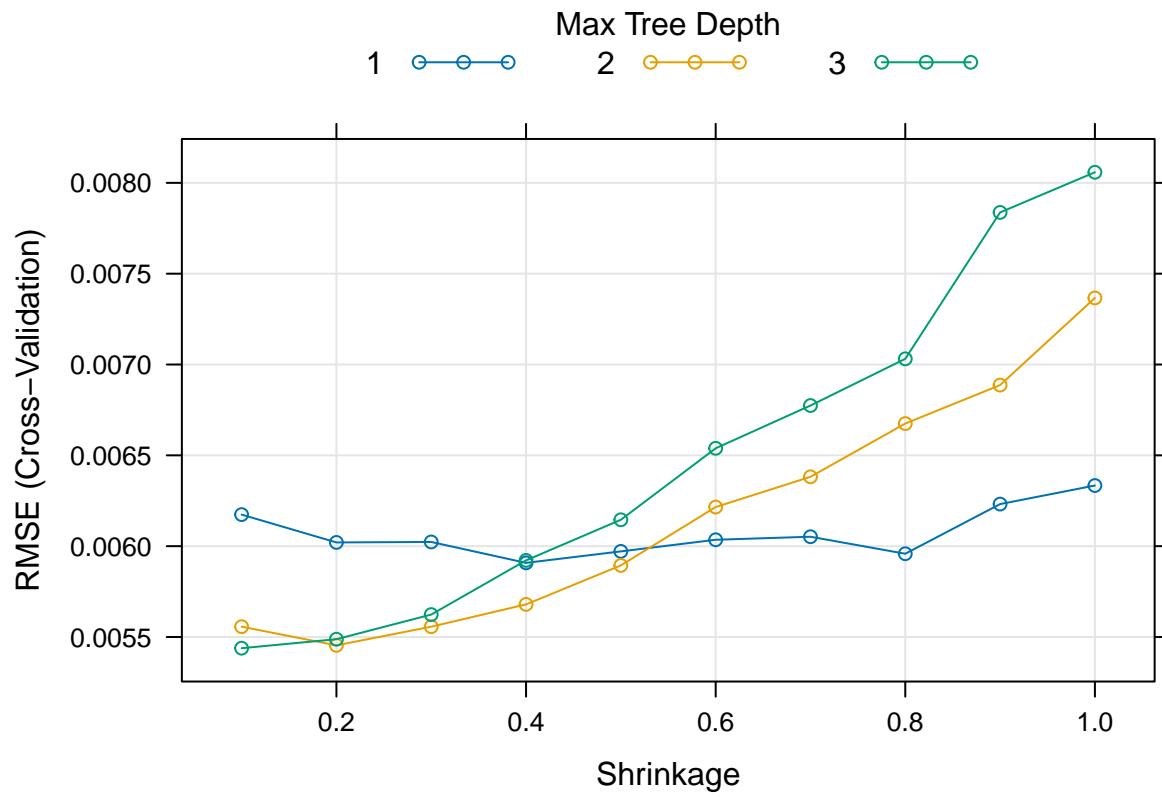
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

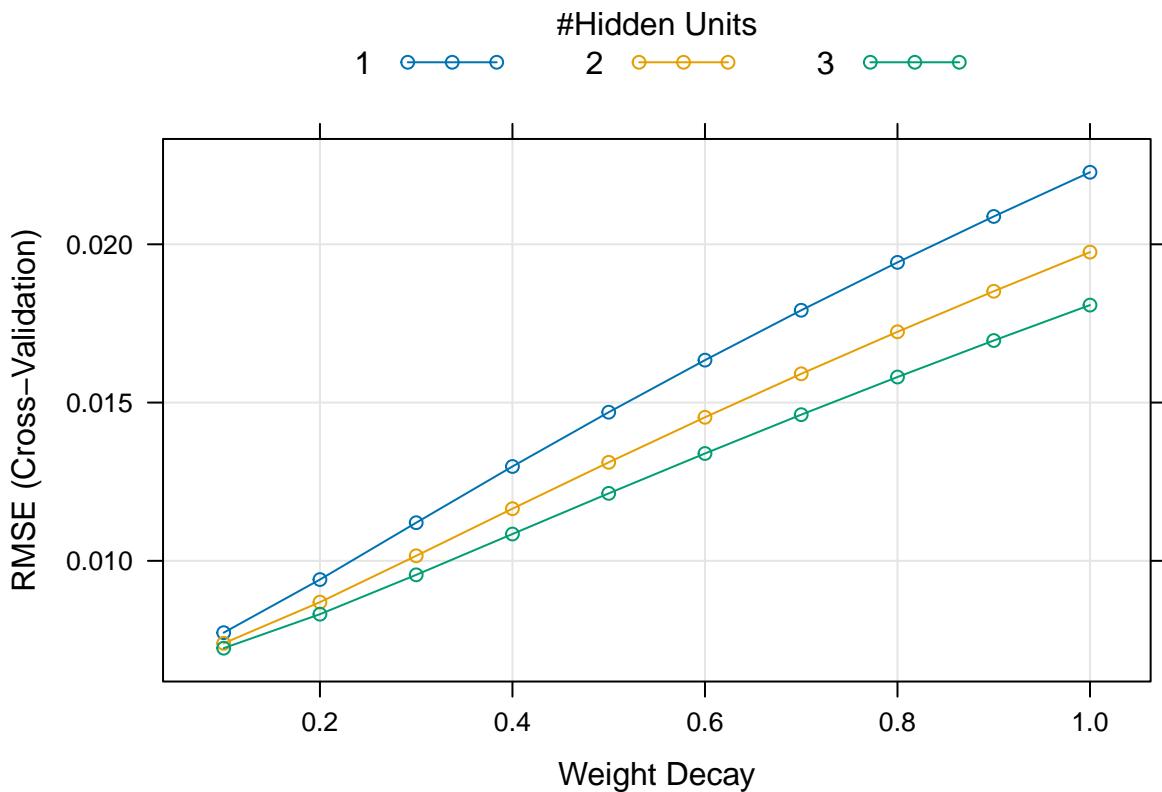
## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.5112323819855e-05"
## [1] "2023-12-09 23:17:20 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.1"
## [1] "2023-12-09 23:17:43 MST"
## [1] "Artificial Neural Net has been fitted"

```

```
## [1] "Best lambda for Artificial Neural Net is: 0.1"  
## [1] "2023-12-09 23:17:47 MST"
```







```

## [1] 1.511232e-05 1.000000e-01 1.000000e-01

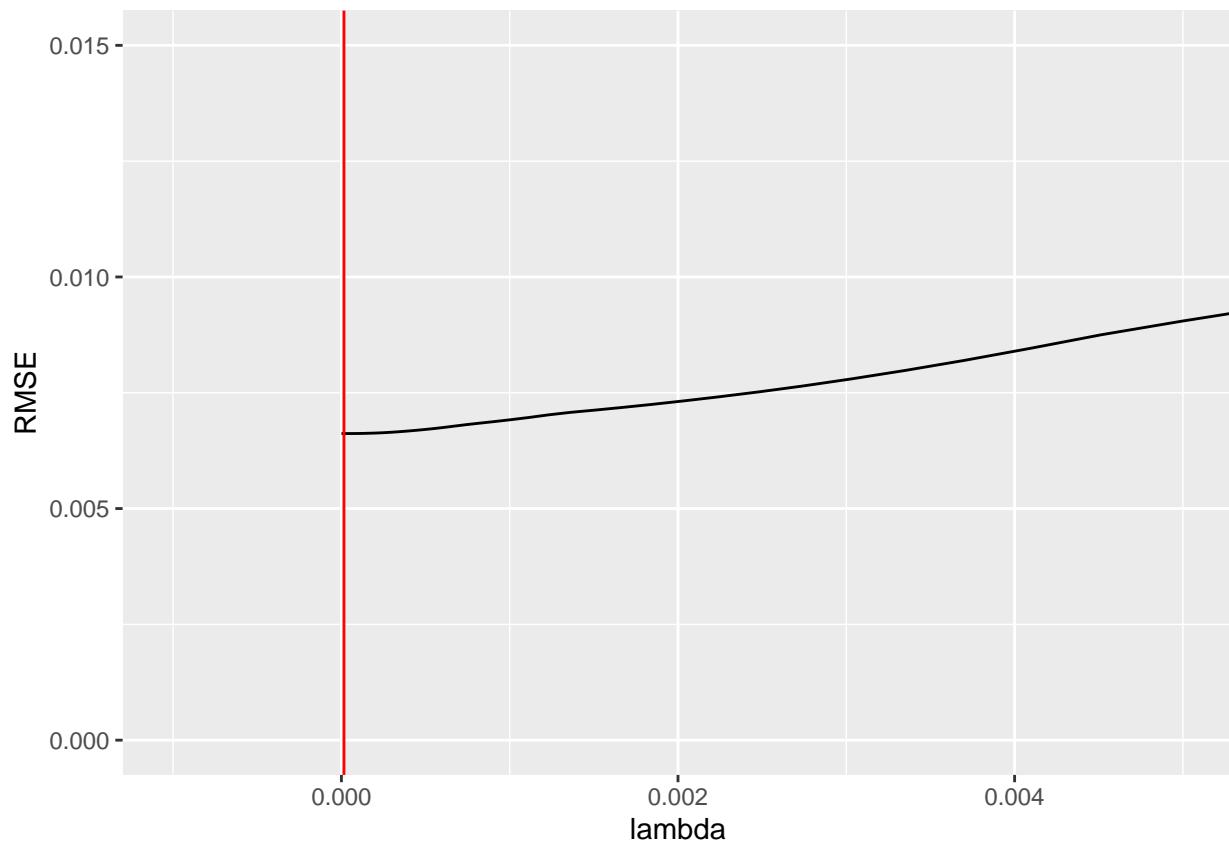
#MAN transformed tuning #3

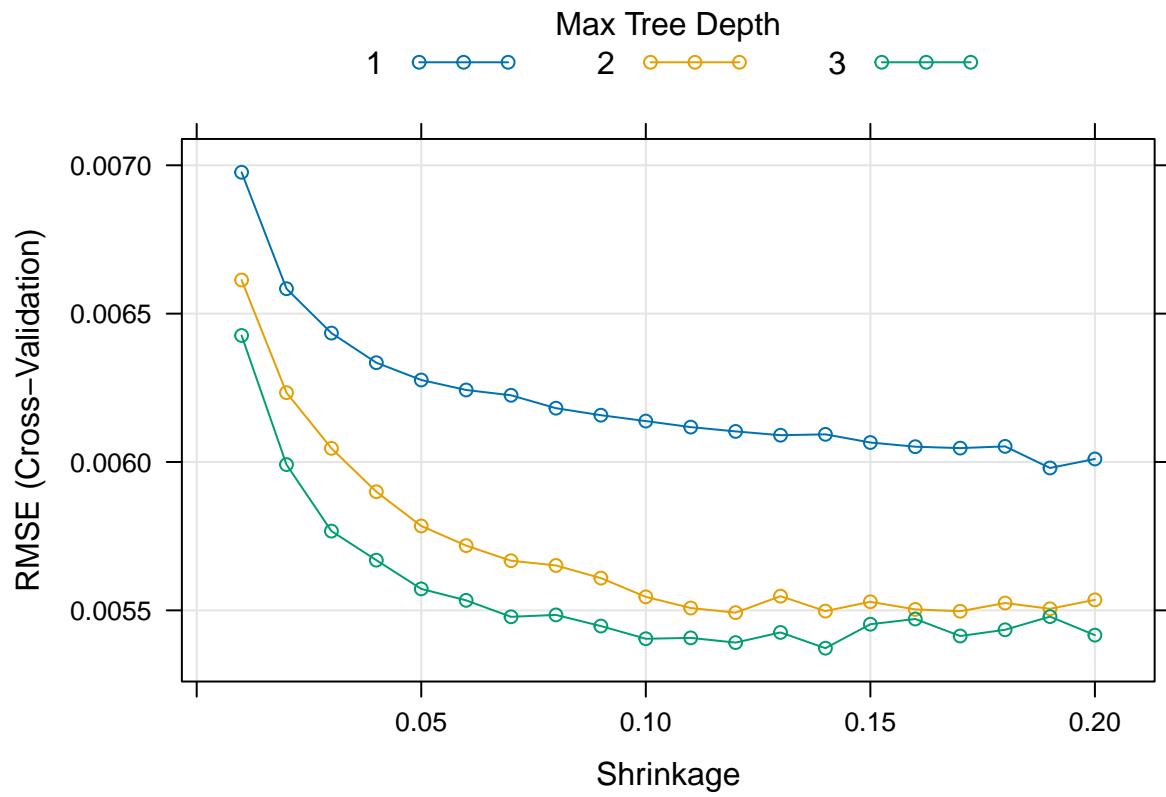
####Updated Values
#Keep LASSO values
boostLambda_MAN_trans = seq(.01,man_trans_tune_2[2]+.1,.01) #best tune is 0.1, lower & upper bound extended
annLambda_MAN_trans = seq(.01,man_trans_tune_2[3],.01) #best tune is 0.1, lower bound extended. Examining the results, the best lambda is 0.14
man_trans_tune_3 = ozone_tune_models(man_8_hr_trans, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean")

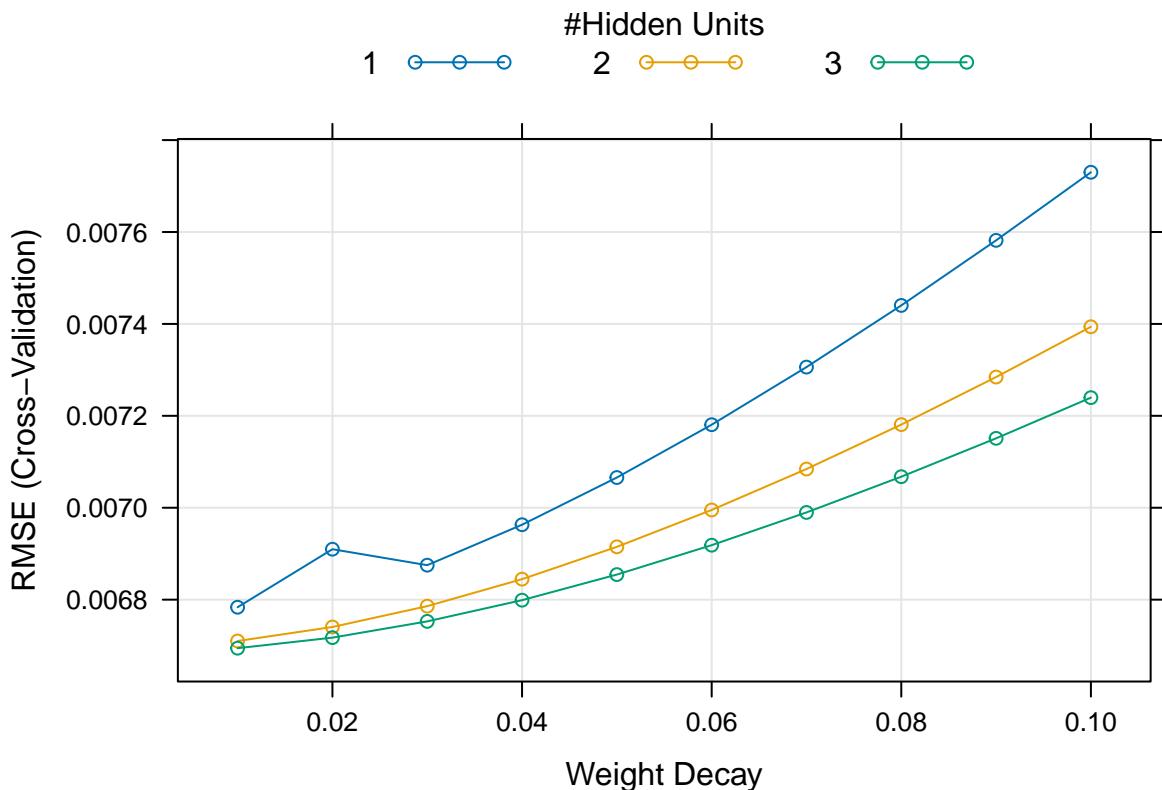
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.5112323819855e-05"
## [1] "2023-12-09 23:17:47 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.14"
## [1] "2023-12-09 23:18:35 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.01"
## [1] "2023-12-09 23:18:39 MST"

```







```

## [1] 1.511232e-05 1.400000e-01 1.000000e-02

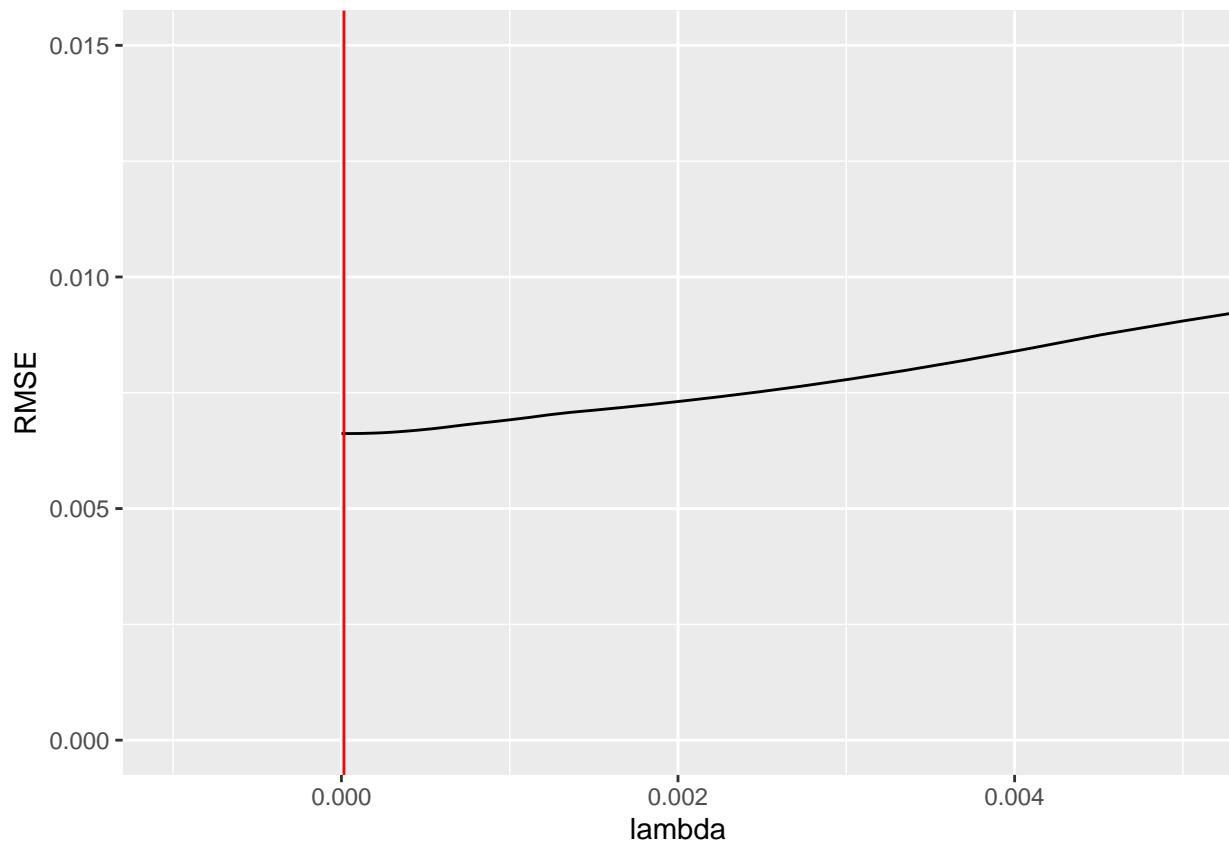
#MAN transformed tuning #4 (final Tune)

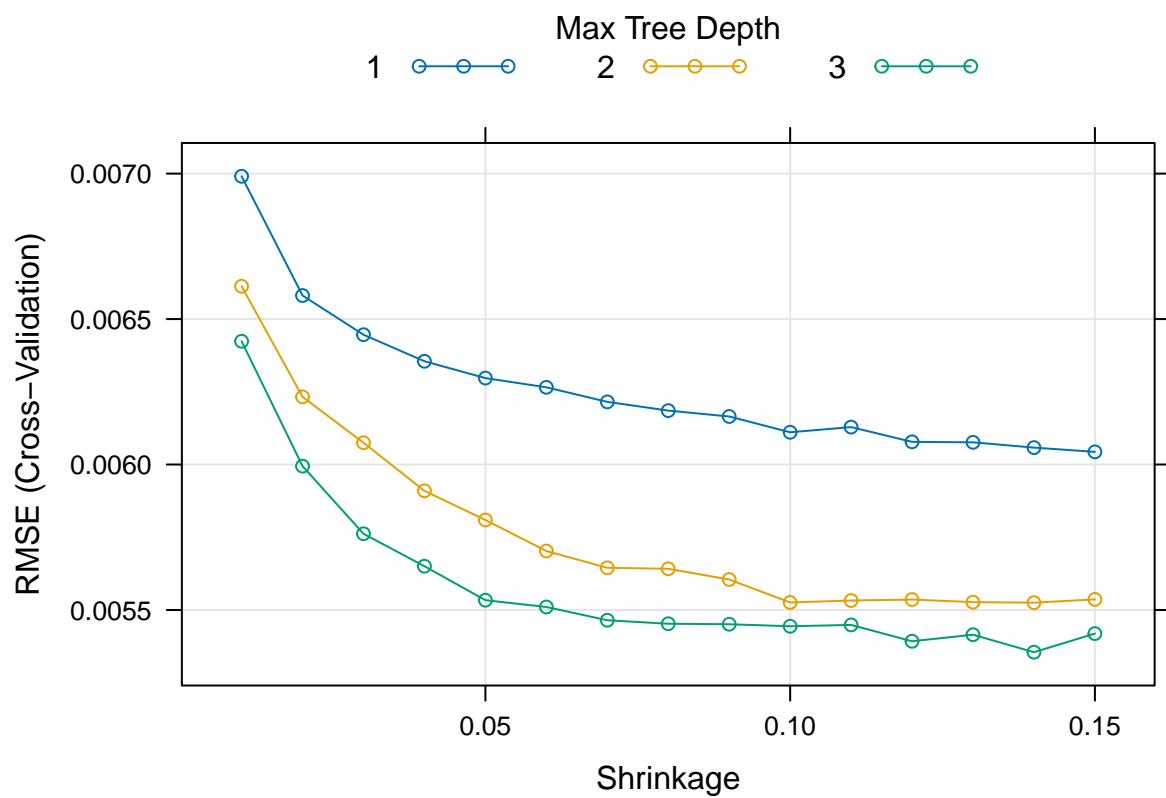
####Updated Values
boostLambda_MAN_trans = seq(.01,man_trans_tune_2[2]+.05,.01) #shorten for final model
annLambda_MAN_trans = seq(.001,man_trans_tune_3[3],.001) #best tune is 0.01, lower bound examined. Example
man_trans_tune_4 = ozone_tune_models(man_8_hr_trans, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean

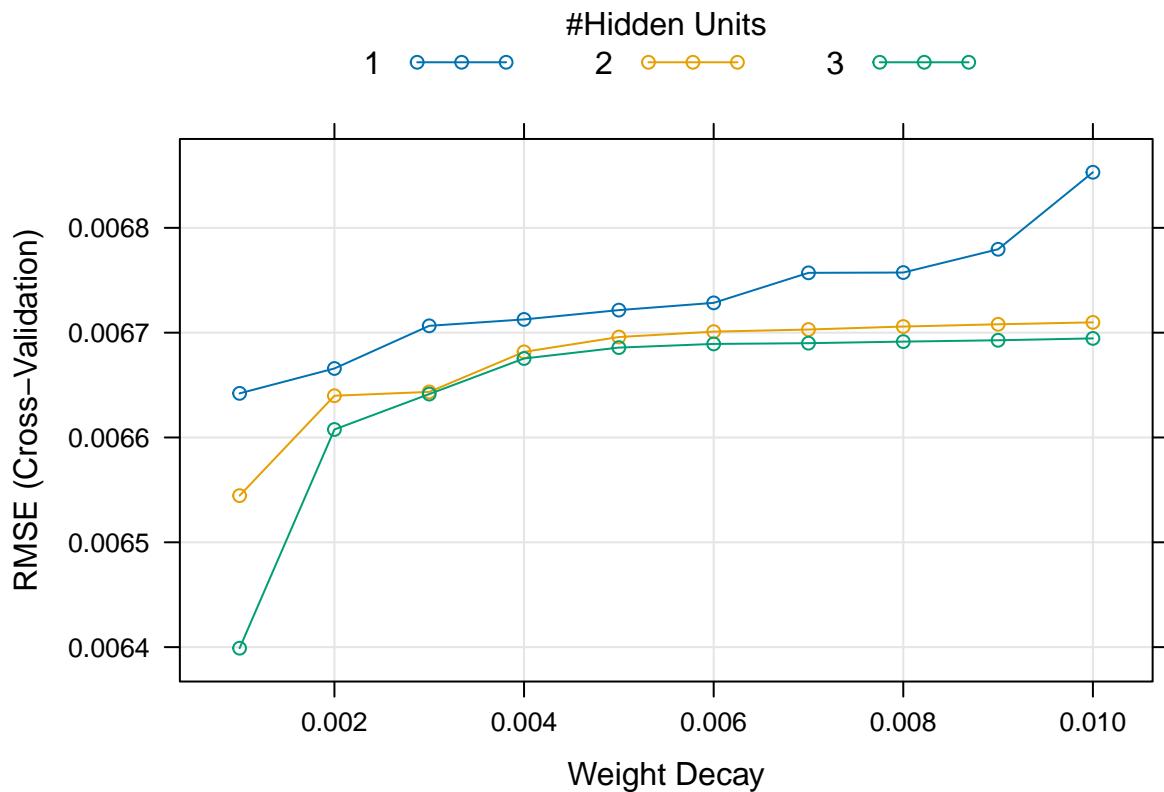
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.

## [1] "LASSO Model has been fitted"
## [1] "Best lambda for LASSO is: 1.5112323819855e-05"
## [1] "2023-12-09 23:18:40 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best lambda for Boosted Trees is: 0.14"
## [1] "2023-12-09 23:19:16 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 0.001"
## [1] "2023-12-09 23:19:23 MST"

```





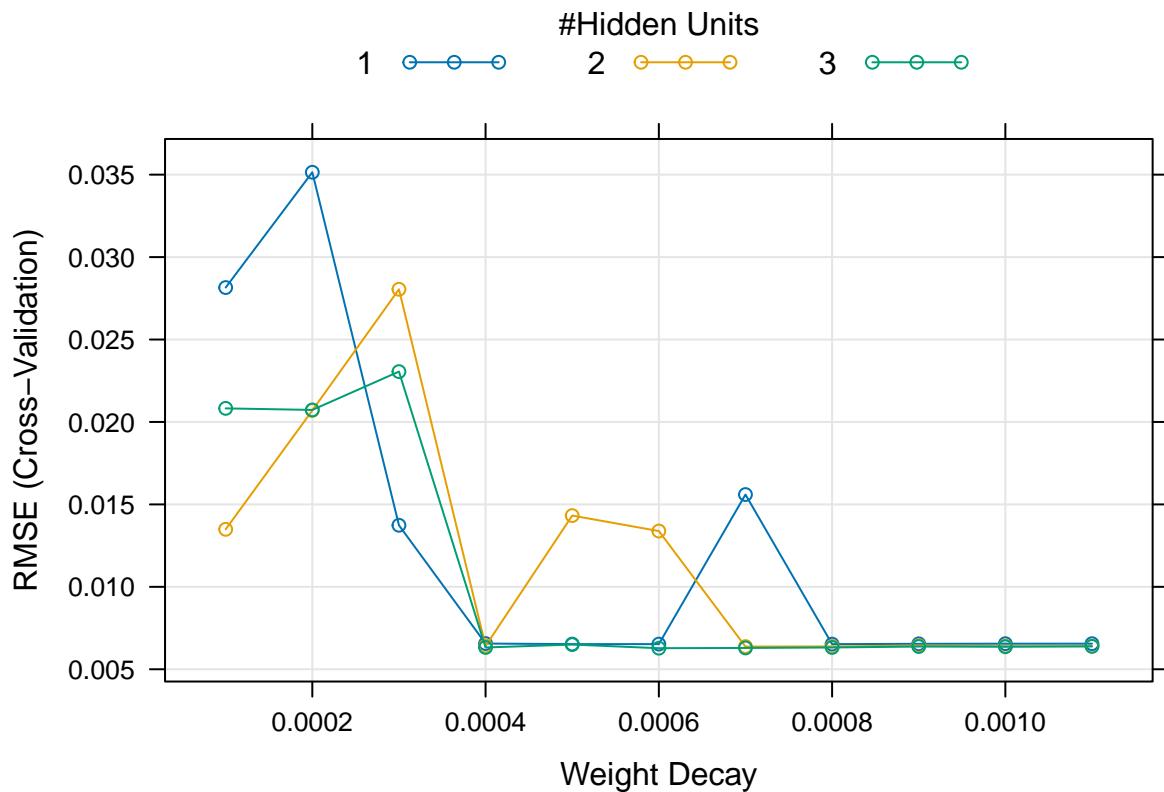


```
## [1] 1.511232e-05 1.400000e-01 1.000000e-03
```

```
#final ann finnessing
annLambda_MAN_trans = seq(.0002,.002,.0002) #best tune is continuously smaller, lower bound extended.
ann = ozone_tuning_ANN(man_8_hr, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",annLambda_MAN)

## [1] "Artificial Neural Net has been fitted"
## [1] "Best lambda for Artificial Neural Net is: 6e-04"
## [1] "2023-12-09 23:19:33 MST"
```

```
plot(ann)
```



PREDICTIVE MODELING

#Model suite function

```

ozone_models <- function(ozone_data_frame, ozone_col, lambdaLASSO, lambdaBOOST, lamdaANN){
  #set up outer cross validation shell
  p = ncol(ozone_data_frame)-1 #Get total number of predictor variables
  n = dim(ozone_data_frame)[1]
  nfolds = 5
  groups = rep(1:nfolds,length=n) #produces list of group labels
  set.seed(28) #for reproducibility
  cvgroups = sample(groups,n) #orders randomly

  # set up storage for predicted values from the double-cross-validation
  allpredictedCV = rep(NA,n)
  # set up storage to see what models are "best" on the inner loops
  allbestTypes = rep(NA,nfolds)
  allbestPars = vector("list",nfolds)
  allbestModels = vector("list",nfolds)
  allbestRSME = rep(NA,nfolds)
  #allbestR2 = rep(NA,nfolds)
  #fold_df_list = vector("list",nfolds)
  all_best_Types = c("Linear", "Generalized Linear Model", "LASSO Regression", "Linear Regression with
  "General Additive Model", "Random Forest", "Gradient Boosted Trees", "Artificial N

```

```

# loop through outer splits
for (j in 1:nfolds) {
  #fold_df = data.frame(all_best_Types)
  fold_msg = paste("Fold #", j)
  print(fold_msg)
  print(Sys.time())
  groupj = (cvgroups == j)
  traindata = ozone_data_frame[!groupj,]
  trainx = model.matrix(traindata[[ozone_col]] ~ ., data = traindata)[,-1]
  trainy = traindata[[ozone_col]]
  validdata = ozone_data_frame[groupj,]
  validx = model.matrix(validdata[[ozone_col]] ~ ., data = validdata)[,-1]
  validy = validdata[[ozone_col]]

  #specify data to be used
  dataused=traindata

  # model-fitting process
  # set up training method
  set.seed(88)
  ozone = dataused[[ozone_col]]
  training = trainControl(method = "cv", number = 5) #5 for computational feasibility
  #fit linear model
  fit_lm = train(formula(paste(ozone_col, " ~ .")),
                 data = dataused,
                 method = "lm",
                 trControl = training)
  print("Linear Regression Model has been fitted")
  r2_message = paste("Best R2: ", round(fit_lm$results$Rsquared, 3))
  print(r2_message)
  print(Sys.time())
  #fit general linear model
  fit_glm = train(formula(paste(ozone_col, " ~ .")),
                  data = dataused,
                  method = "glm",
                  trControl = training)
  print("Generalized Linear Model has been fitted")#debug
  r2_message = paste("Best R2: ", round(fit_glm$results$Rsquared, 3))
  print(r2_message)
  print(Sys.time())
  #fit linear regression with LASSO
  fit_LASSO = train(formula(paste(ozone_col, " ~ .")),
                    data = dataused,
                    method = "glmnet",
                    trControl = training,
                    tuneGrid = expand.grid(alpha=c(1), lambda=lambdaLASSO))
  print("LASSO Model has been fitted")#debug
  r2_message = paste("Best R2: ", round(max(fit_LASSO$results$Rsquared, na.rm=T), 3))
  print(r2_message)
  print(Sys.time())
  #fit linear regression with forward selection model
  fit_fs = train(formula(paste(ozone_col, " ~ .")),
                 data = dataused,

```

```

        method = "leapForward",
        trControl = training,
        tuneGrid = expand.grid(nvmax = seq(1, p, 1))) #p is max number of variables w
print("Linear Regression with Forward Selection Model has been fitted")
r2_message = paste("Best R2: ", round(max(fit_fs$results$Rsquared), 3))
print(r2_message)
print(Sys.time())
#fit GAM with splines
fit_gam_spline = train(formula(paste(ozone_col, " ~ .")),
                       data = dataused,
                       method = "gam",
                       trControl = training,
                       tuneGrid = data.frame(method="REML", select=TRUE))
print("GAM with Splines has been fitted") #debug
r2_message = paste("Best R2: ", round(max(fit_gam_spline$results$Rsquared), 3))
print(r2_message)
print(Sys.time())
#fit Random Forest model
fit_rf = train(formula(paste(ozone_col, " ~ .")),
               data = dataused,
               method = "rf",
               tuneGrid = expand.grid(mtry = 1:5),
               trControl = training)
print("Random Forest has been fitted") #debug
r2_message = paste("Best R2: ", round(max(fit_rf$results$Rsquared), 3))
print(r2_message)
print(Sys.time())
#fit Boosted Trees
fit_boost = train(formula(paste(ozone_col, " ~ .")),
                  data = dataused,
                  method = "gbm",
                  trControl = training,
                  tuneGrid = expand.grid(interaction.depth = (1:3), #use smaller interaction
                                         n.trees = 300, #use 300 trees (more trees is ideal, but comp
                                         shrinkage = lambdaBOOST, # tuned hyperparameter vector
                                         n.minobsinnode = 10), # use default
                                         verbose=FALSE) #no readout
print("Boosted Trees has been fitted") #debug
r2_message = paste("Best R2: ", round(max(fit_boost$results$Rsquared), 3))
print(r2_message)
print(Sys.time())
#fit Artificial Neural Net
fit ANN = train(formula(paste(ozone_col, " ~ .")),
                 data = dataused,
                 method = "nnet",
                 tuneGrid = expand.grid(size = (1:3), decay = lamdaANN), #may need to limit for compu
                 trace = FALSE,
                 preProc = c("center", "scale"), #good idea to prevent one predictor from overwhelmin
                 trControl = training,
                 verbose=FALSE)
print("Artificial Neural Net has been fitted") #debug
r2_message = paste("Best R2: ", round(max(fit ANN$results$Rsquared, na.rm = T), 3))
print(r2_message)

```

```

### identify selected model to fit to full data
all_best_Parms = list(fit_lm$bestTune,fit_glm$bestTune,fit_LASSO$bestTune,fit_fs$bestTune,fit_gam_spline$bestTune,
                      fit_boost$bestTune,fit_ANN$bestTune)
all_best_Models = list(fit_lm$finalModel,fit_glm,fit_LASSO,fit_fs,fit_gam_spline,fit_rf,
                       fit_boost$finalModel,fit_ANN)
all_best_RMSE = c(fit_lm$results$RMSE,fit_glm$results$RMSE,min(fit_LASSO$results$RMSE),min(fit_fs$results$RMSE),
                  fit_gam_spline$results$RMSE,min(fit_rf$results$RMSE),min(fit_boost$results$RMSE),),
one_best_Type = all_best_Types[which.min(all_best_RMSE)]
one_best_Parms = all_best_Parms[which.min(all_best_RMSE)]
one_best_Model = all_best_Models[which.min(all_best_RMSE)]

# store
allbestTypes[j] = one_best_Type
allbestParms[[j]] = one_best_Parms
#store to compare models
allbestModels[[j]] = one_best_Model
allbestRSME[j] = min(all_best_RMSE)
#print("To the loops!") #debug

if (one_best_Type == "Linear") { #linear model
  print("----->Linear model works best")
  ###one_best_Model not fitting correctly
  allpredictedCV[groupj] = predict(lm(dataused[[ozone_col]] ~ ., data=dataused),validdata) ### Predicted values

} else if (one_best_Type == "Generalized Linear Model") {
  print("----->General Linear Model works best")
  allpredictedCV[groupj] = predict(one_best_Model,validdata)

} else if (one_best_Type == "LASSO Regression") {
  print("----->LASSO Regression works best")
  allpredictedCV[groupj] = predict.train(fit_LASSO,validdata)

} else if (one_best_Type == "Linear Regression with Forward Selection") { #linear with selection
  print("----->Forward Regression works best")
  allpredictedCV[groupj] = predict.train(fit_fs,validdata)

} else if (one_best_Type == "General Additive Model") { ###GAM with p-splines
  print("----->General Additive Model works best")
  allpredictedCV[groupj] = predict.train(fit_gam_spline,validdata)

} else if (one_best_Type == "Random Forest") {
  print("----->Random Forest works best")
  allpredictedCV[groupj] = predict.train(fit_rf,validdata,type="raw") #raw for numeric predictions

} else if (one_best_Type == "Gradient Boosted Trees") {
  print("----->Gradient Boosted Trees works best")
  allpredictedCV[groupj] = predict.train(fit_boost,validdata) ### should I use predict() or predict.trees()

} else if (one_best_Type == "Artificial Neural Net") {
  print("----->Artificial Neural Net works best")
  allpredictedCV[groupj] = predict.train(fit_ANN,validdata) ### should I use predict() or predict.trees()
  ### why doesn't it work with one_best_Model
}

}

```

```

}

#assessment
y = ozone_data_frame[[ozone_col]]
RMSE = sqrt(mean(allpredictedCV-y)^2)
R2 = 1-sum((allpredictedCV-y)^2)/sum((y-mean(y))^2)
final_model = allbestModels[[which.min(allbestRSME)]]
final_type = allbestTypes[[which.min(allbestRSME)]]
final_pars = allbestPars[[which.min(allbestRSME)]]]

res_list = c(final_model, final_type, final_pars)

rmse = paste("Final RSME:", RMSE)
r2 = paste("Final R2:", R2)
finMod = paste("Overall Best Model:", final_type)
print(rmse)
print(r2)
print(finMod)
print(Sys.time())
return(res_list)
} # end function ozone_models

```

#AFA 8 Hour

```

afa_8_hr = ozone_8_hr[,-c(2)]
afa_8_hr_cv_results = ozone_models(afa_8_hr, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",
                                    lassoLambda_AFA, boostLambda_AFA, annLambda_AFA)

## [1] "Fold # 1"
## [1] "2023-12-09 23:19:33 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.525"
## [1] "2023-12-09 23:19:33 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.52"
## [1] "2023-12-09 23:19:33 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.523"
## [1] "2023-12-09 23:19:34 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.523"
## [1] "2023-12-09 23:19:34 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.541"
## [1] "2023-12-09 23:21:39 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.603"
## [1] "2023-12-09 23:22:30 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.655"
## [1] "2023-12-09 23:22:56 MST"
## [1] "Artificial Neural Net has been fitted"

```

```

## [1] "Best R2: 0.537"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 2"
## [1] "2023-12-09 23:23:03 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.529"
## [1] "2023-12-09 23:23:04 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.532"
## [1] "2023-12-09 23:23:04 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.529"
## [1] "2023-12-09 23:23:04 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.531"
## [1] "2023-12-09 23:23:04 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.545"
## [1] "2023-12-09 23:24:53 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.602"
## [1] "2023-12-09 23:25:45 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.647"
## [1] "2023-12-09 23:26:10 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.539"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 3"
## [1] "2023-12-09 23:26:17 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.52"
## [1] "2023-12-09 23:26:17 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.518"
## [1] "2023-12-09 23:26:17 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.519"
## [1] "2023-12-09 23:26:17 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.519"
## [1] "2023-12-09 23:26:17 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.534"
## [1] "2023-12-09 23:28:15 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.592"
## [1] "2023-12-09 23:29:06 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.653"
## [1] "2023-12-09 23:29:31 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.536"
## [1] "----->Gradient Boosted Trees works best"

```

```

## [1] "Fold # 4"
## [1] "2023-12-09 23:29:38 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.518"
## [1] "2023-12-09 23:29:39 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.519"
## [1] "2023-12-09 23:29:39 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.518"
## [1] "2023-12-09 23:29:39 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.52"
## [1] "2023-12-09 23:29:39 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.526"
## [1] "2023-12-09 23:31:41 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.596"
## [1] "2023-12-09 23:32:33 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.629"
## [1] "2023-12-09 23:32:58 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.526"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 5"
## [1] "2023-12-09 23:33:05 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.524"
## [1] "2023-12-09 23:33:05 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.52"
## [1] "2023-12-09 23:33:05 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.523"
## [1] "2023-12-09 23:33:05 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.521"
## [1] "2023-12-09 23:33:06 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.544"
## [1] "2023-12-09 23:35:24 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.6"
## [1] "2023-12-09 23:36:16 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.645"
## [1] "2023-12-09 23:36:41 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.528"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Final RSME: 3.14330092811016e-05"
## [1] "Final R2: 0.648082297045422"

```

```

## [1] "Overall Best Model: Gradient Boosted Trees"
## [1] "2023-12-09 23:36:48 MST"

#AFA Final Fit

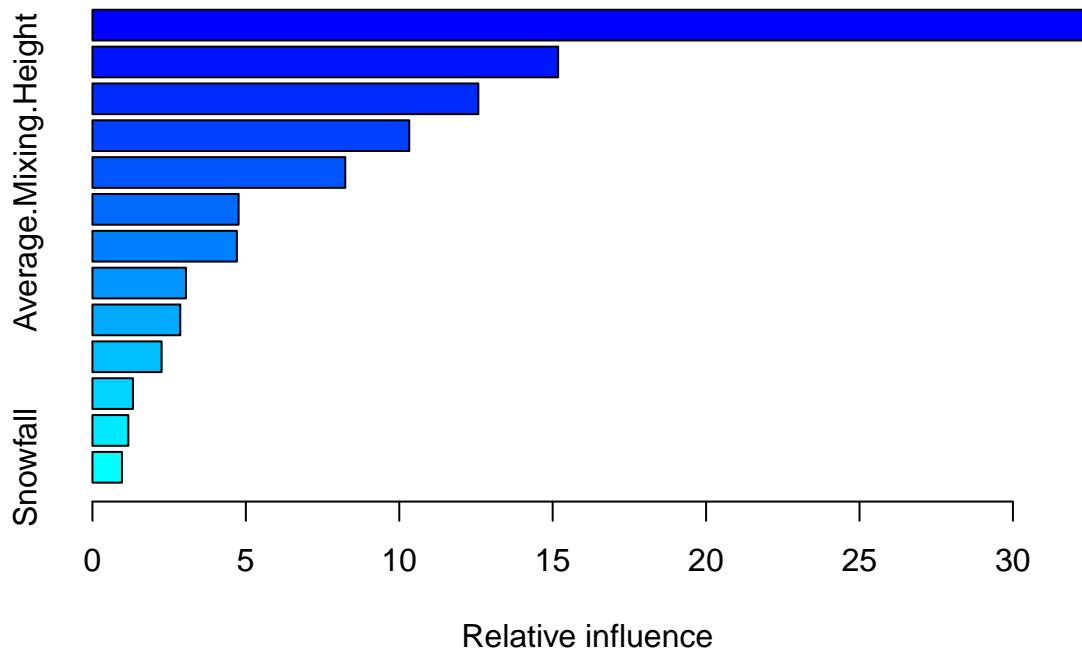
afa_final_pars = data.frame(afa_8_hr_cv_results[3])
training = trainControl(method = "cv", number = 10)
afa_finalModelGrid <- data.frame(interaction.depth = afa_final_pars$interaction.depth,
                                   n.trees = afa_final_pars$n.trees,
                                   shrinkage = afa_final_pars$shrinkage,
                                   n.minobsinnode = afa_final_pars$n.minobsinnode) # use default

# cross-validation of Boosted Tree

afa_fit_final = train(AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~.,
                      data = afa_8_hr,
                      verbose=FALSE,
                      method = "gbm",
                      trControl = training,
                      tuneGrid = afa_finalModelGrid)

summary(afa_fit_final)

```



```

##
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean

```

```

HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean

```

```

## Date HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
## Fastest.2.minute.wind.speed Fastest.2.minute.wind.speed
## Average.Mixing.Height Average.Mixing.Height
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
## Direction.of.fastest.2.minute.wind Direction.of.fastest.2.minute.wind
## Snow.depth
## Precipitation
## Snowfall
## rel.inf
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean 32.5891029
## Date 15.1779493
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean 12.5769072
## Fastest.2.minute.wind.speed 10.3256592
## Average.Mixing.Height 8.2405598
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 4.7647504
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean 4.7069591
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 3.0480237
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean 2.8631704
## Direction.of.fastest.2.minute.wind 2.2546282
## Snow.depth 1.3205780
## Precipitation 1.1694869
## Snowfall 0.9622248

afa_fit_final

## Stochastic Gradient Boosting
##
## 2017 samples
##    13 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1815, 1817, 1815, 1814, 1815, 1815, ...
## Resampling results:
##
##      RMSE      Rsquared     MAE
##      0.005964741  0.657107  0.004751047
##
## Tuning parameter 'n.trees' was held constant at a value of 300
## Tuning
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.1
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10

#AFA Transformed

afa_transformed = ozone_transformed[-c(2)]
afa_trans_cv_results = ozone_models(afa_transformed, "AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
                                    lassoLambda_AFA_trans, boostLambda_AFA_trans, annLambda_AFA_trans)

```

```

## [1] "Fold # 1"
## [1] "2023-12-09 23:36:52 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.507"
## [1] "2023-12-09 23:36:52 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.506"
## [1] "2023-12-09 23:36:52 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.511"
## [1] "2023-12-09 23:36:52 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.508"
## [1] "2023-12-09 23:36:53 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.547"
## [1] "2023-12-09 23:38:30 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.597"
## [1] "2023-12-09 23:39:28 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.65"
## [1] "2023-12-09 23:39:58 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.526"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 2"
## [1] "2023-12-09 23:40:04 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.509"
## [1] "2023-12-09 23:40:04 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.513"
## [1] "2023-12-09 23:40:04 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.511"
## [1] "2023-12-09 23:40:05 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.513"
## [1] "2023-12-09 23:40:05 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.547"
## [1] "2023-12-09 23:41:55 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.598"
## [1] "2023-12-09 23:42:54 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.629"
## [1] "2023-12-09 23:43:23 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.524"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 3"
## [1] "2023-12-09 23:43:29 MST"

```

```

## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.503"
## [1] "2023-12-09 23:43:29 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.501"
## [1] "2023-12-09 23:43:29 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.502"
## [1] "2023-12-09 23:43:29 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.503"
## [1] "2023-12-09 23:43:30 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.535"
## [1] "2023-12-09 23:45:05 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.585"
## [1] "2023-12-09 23:46:02 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.64"
## [1] "2023-12-09 23:46:31 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.514"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 4"
## [1] "2023-12-09 23:46:37 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.497"
## [1] "2023-12-09 23:46:37 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.499"
## [1] "2023-12-09 23:46:37 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.499"
## [1] "2023-12-09 23:46:38 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.502"
## [1] "2023-12-09 23:46:38 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.529"
## [1] "2023-12-09 23:48:41 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.593"
## [1] "2023-12-09 23:49:40 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.621"
## [1] "2023-12-09 23:50:09 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.512"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 5"
## [1] "2023-12-09 23:50:16 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.502"

```

```

## [1] "2023-12-09 23:50:16 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.5"
## [1] "2023-12-09 23:50:16 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.501"
## [1] "2023-12-09 23:50:16 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.501"
## [1] "2023-12-09 23:50:16 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.54"
## [1] "2023-12-09 23:52:03 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.593"
## [1] "2023-12-09 23:52:59 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.637"
## [1] "2023-12-09 23:53:29 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.515"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Final RSME: 4.67329015170942e-05"
## [1] "Final R2: 0.640722089544293"
## [1] "Overall Best Model: Gradient Boosted Trees"
## [1] "2023-12-09 23:53:35 MST"

```

#Fit final model

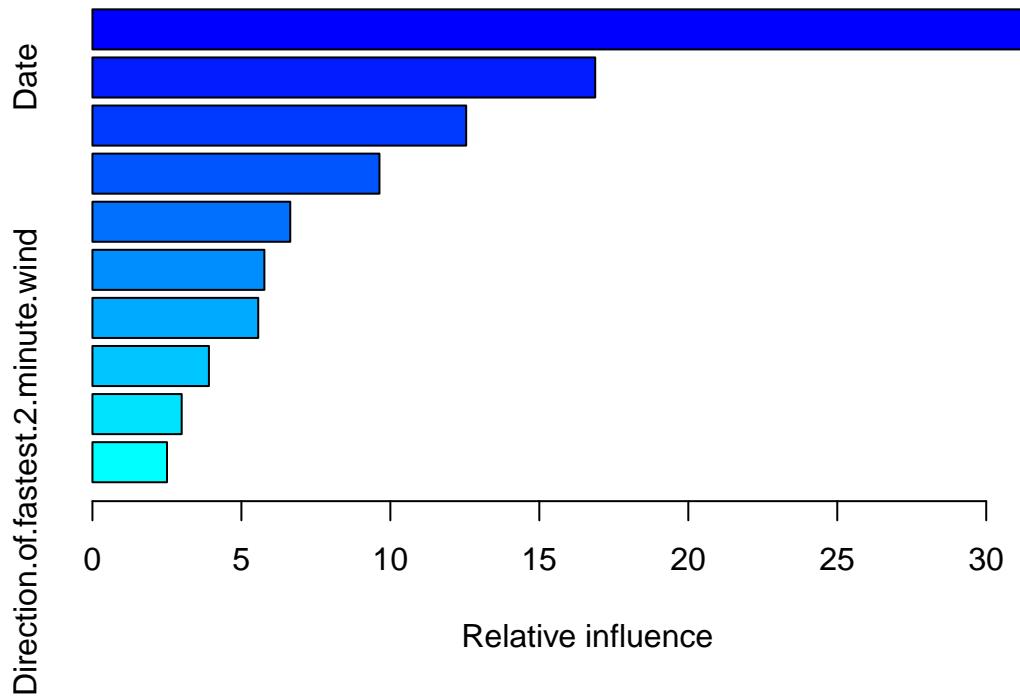
```

afa_trans_final_pars = data.frame(afa_trans_cv_results[3])
training = trainControl(method = "cv", number = 10)
afa_trans_finalModelGrid <- data.frame(interaction.depth = afa_trans_final_pars$interaction.depth,
                                         n.trees = afa_trans_final_pars$n.trees,
                                         shrinkage = afa_trans_final_pars$shrinkage,
                                         n.minobsinnode = afa_trans_final_pars$n.minobsinnode) # use default

# cross-validation of Boosted Tree
afa_trans_fit_final = train(AFA_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~ .,
                            data = afa_transformed,
                            verbose=FALSE,
                            method = "gbm",
                            trControl = training,
                            tuneGrid = afa_trans_finalModelGrid)

summary(afa_trans_fit_final)

```



```

## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
## Date HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## Fastest.2.minute.wind.speed HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## Average.Mixing.Height.power HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean
## log_HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean log_HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
## Direction.of.fastest.2.minute.wind Direction.of.fastest.2.minute.wind
## rel.inf
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean 33.562502
## Date 16.875260
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean 12.544015
## Fastest.2.minute.wind.speed 9.633455
## Average.Mixing.Height.power 6.639725
## log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 5.771690
## log_HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean 5.564657
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 3.909948
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean 2.995859
## Direction.of.fastest.2.minute.wind 2.502889

afa_trans_fit_final

## Stochastic Gradient Boosting

```

```

## 
## 2017 samples
##   10 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1813, 1815, 1814, 1817, 1815, 1815, ...
## Resampling results:
##
##    RMSE      Rsquared     MAE
##    0.006051406  0.6465636  0.004801719
##
## Tuning parameter 'n.trees' was held constant at a value of 300
## Tuning
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.13
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10

```

#MAN 8 Hour

```

man_8_hr = ozone_8_hr[-c(1)]
man_8_hr_cv_results = ozone_models(man_8_hr, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean",
                                    lassoLambda_MAN, boostLambda_MAN, annLambda_MAN)

```

```

## [1] "Fold # 1"
## [1] "2023-12-09 23:53:38 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.549"
## [1] "2023-12-09 23:53:38 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.547"
## [1] "2023-12-09 23:53:38 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.547"
## [1] "2023-12-09 23:53:38 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.545"
## [1] "2023-12-09 23:53:38 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.571"
## [1] "2023-12-09 23:56:19 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.644"
## [1] "2023-12-09 23:57:11 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.695"
## [1] "2023-12-09 23:57:36 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.586"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 2"
## [1] "2023-12-09 23:57:44 MST"

```

```

## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.552"
## [1] "2023-12-09 23:57:44 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.553"
## [1] "2023-12-09 23:57:44 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.552"
## [1] "2023-12-09 23:57:44 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.554"
## [1] "2023-12-09 23:57:45 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.569"
## [1] "2023-12-09 23:59:47 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.65"
## [1] "2023-12-10 00:00:40 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.682"
## [1] "2023-12-10 00:01:05 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.582"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 3"
## [1] "2023-12-10 00:01:12 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.545"
## [1] "2023-12-10 00:01:13 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.544"
## [1] "2023-12-10 00:01:13 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.544"
## [1] "2023-12-10 00:01:13 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.544"
## [1] "2023-12-10 00:01:13 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.562"
## [1] "2023-12-10 00:03:28 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.641"
## [1] "2023-12-10 00:04:20 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.697"
## [1] "2023-12-10 00:04:45 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.58"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 4"
## [1] "2023-12-10 00:04:53 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.541"

```

```

## [1] "2023-12-10 00:04:53 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.541"
## [1] "2023-12-10 00:04:53 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.543"
## [1] "2023-12-10 00:04:53 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.546"
## [1] "2023-12-10 00:04:53 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.557"
## [1] "2023-12-10 00:07:08 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.633"
## [1] "2023-12-10 00:07:59 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.692"
## [1] "2023-12-10 00:08:23 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.564"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 5"
## [1] "2023-12-10 00:08:31 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.545"
## [1] "2023-12-10 00:08:31 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.546"
## [1] "2023-12-10 00:08:31 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.542"
## [1] "2023-12-10 00:08:32 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.545"
## [1] "2023-12-10 00:08:32 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.562"
## [1] "2023-12-10 00:10:41 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.642"
## [1] "2023-12-10 00:11:32 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.692"
## [1] "2023-12-10 00:11:57 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.583"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Final RSME: 1.98747566244226e-05"
## [1] "Final R2: 0.693723906142193"
## [1] "Overall Best Model: Gradient Boosted Trees"
## [1] "2023-12-10 00:12:05 MST"

```

#Fit MAN 8 Hour Final Model

```

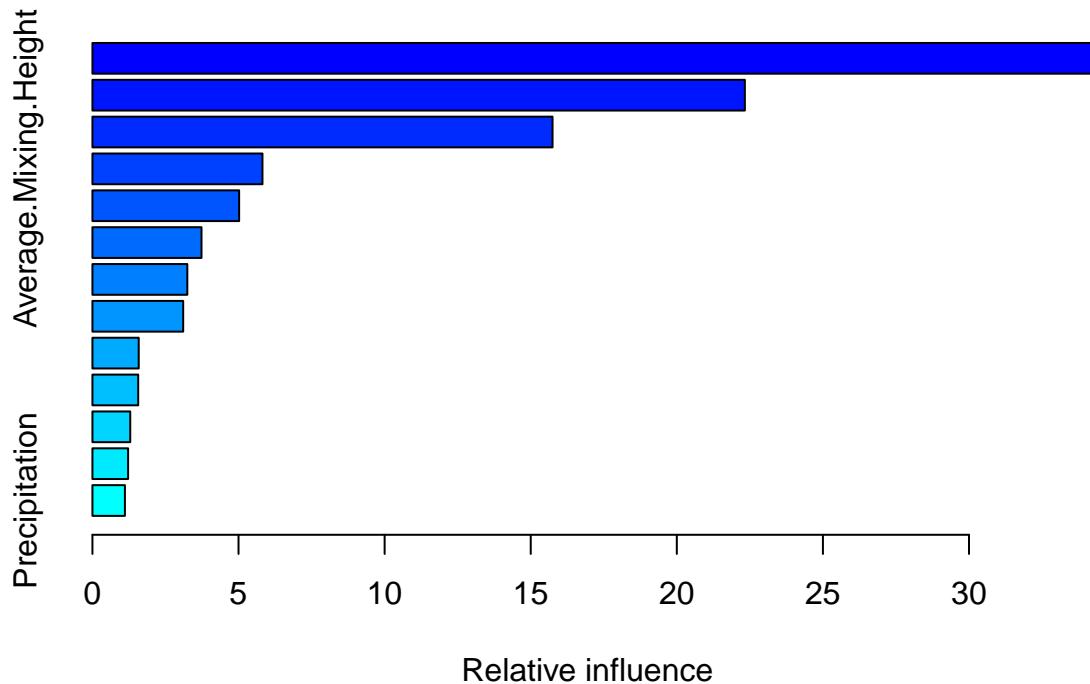
man_final_pars = data.frame(man_8_hr_cv_results[3])
training = trainControl(method = "cv", number = 10)
man_finalModelGrid <- data.frame(interaction.depth = man_final_pars$interaction.depth,
                                   n.trees = man_final_pars$n.trees,
                                   shrinkage = man_final_pars$shrinkage,
                                   n.minobsinnode = man_final_pars$n.minobsinnode) # use default

# cross-validation of Boosted Tree

man_fit_final = train(MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~.,
                      data = man_8_hr,
                      verbose=FALSE,
                      method = "gbm",
                      trControl = training,
                      tuneGrid = man_finalModelGrid)

summary(man_fit_final)

```



```

## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean          HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean          HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
## Date
## Average.Mixing.Height
## HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean          HW24_S02.max.5.min.avg_1.HOUR_Arithmetic.Mean
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean

```

```

## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## Snow.depth
## Fastest.2.minute.wind.speed
## Direction.of.fastest.2.minute.wind
## Snowfall
## Precipitation
##
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
## Date
## Average.Mixing.Height
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## Snow.depth
## Fastest.2.minute.wind.speed
## Direction.of.fastest.2.minute.wind
## Snowfall
## Precipitation

```

HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean	Fastest.2.minute.wind.speed
HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean	Direction.of.fastest.2.minute.wind
## Snow.depth	
## Precipitation	
##	rel.inf
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean	34.221854
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean	22.328834
## Date	15.746790
## Average.Mixing.Height	5.818712
## HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean	5.020135
## HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean	3.733167
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean	3.247660
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean	3.105744
## Snow.depth	1.584503
## Fastest.2.minute.wind.speed	1.566406
## Direction.of.fastest.2.minute.wind	1.294473
## Snowfall	1.220196
## Precipitation	1.111527

```
man_fit_final
```

```

## Stochastic Gradient Boosting
##
## 2017 samples
##    13 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 1814, 1816, 1814, 1814, 1814, 1817, ...
## Resampling results:
##
##     RMSE      Rsquared      MAE
## 0.005230584  0.7078429  0.004119809
##
## Tuning parameter 'n.trees' was held constant at a value of 300
## Tuning
##
## Tuning parameter 'shrinkage' was held constant at a value of 0.13
##
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10

```

```
#MAN Transformed
```

```

man_transformed = ozone_transformed[-c(1)]
man_trans_cv_results = ozone_models(man_transformed, "MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean"
                                         lassoLambda_MAN_trans, boostLambda_MAN_trans, annLambda_MAN_trans)

```

```

## [1] "Fold # 1"
## [1] "2023-12-10 00:12:08 MST"
## [1] "Linear Regression Model has been fitted"

```

```

## [1] "Best R2: 0.532"
## [1] "2023-12-10 00:12:08 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.531"
## [1] "2023-12-10 00:12:08 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.532"
## [1] "2023-12-10 00:12:09 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.53"
## [1] "2023-12-10 00:12:09 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.569"
## [1] "2023-12-10 00:13:59 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.642"
## [1] "2023-12-10 00:14:58 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.673"
## [1] "2023-12-10 00:15:27 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.553"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 2"
## [1] "2023-12-10 00:15:33 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.539"
## [1] "2023-12-10 00:15:33 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.54"
## [1] "2023-12-10 00:15:33 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.537"
## [1] "2023-12-10 00:15:34 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.539"
## [1] "2023-12-10 00:15:34 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.572"
## [1] "2023-12-10 00:17:37 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.642"
## [1] "2023-12-10 00:18:36 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.678"
## [1] "2023-12-10 00:19:05 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.562"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 3"
## [1] "2023-12-10 00:19:11 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.527"
## [1] "2023-12-10 00:19:11 MST"

```

```

## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.528"
## [1] "2023-12-10 00:19:11 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.528"
## [1] "2023-12-10 00:19:11 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.528"
## [1] "2023-12-10 00:19:12 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.565"
## [1] "2023-12-10 00:21:16 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.632"
## [1] "2023-12-10 00:22:14 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.688"
## [1] "2023-12-10 00:22:43 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.563"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 4"
## [1] "2023-12-10 00:22:49 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.529"
## [1] "2023-12-10 00:22:49 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.53"
## [1] "2023-12-10 00:22:49 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.532"
## [1] "2023-12-10 00:22:50 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.534"
## [1] "2023-12-10 00:22:50 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.556"
## [1] "2023-12-10 00:24:41 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.628"
## [1] "2023-12-10 00:25:39 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.682"
## [1] "2023-12-10 00:26:08 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.558"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Fold # 5"
## [1] "2023-12-10 00:26:14 MST"
## [1] "Linear Regression Model has been fitted"
## [1] "Best R2: 0.529"
## [1] "2023-12-10 00:26:14 MST"
## [1] "Generalized Linear Model has been fitted"
## [1] "Best R2: 0.531"

```

```

## [1] "2023-12-10 00:26:14 MST"
## [1] "LASSO Model has been fitted"
## [1] "Best R2: 0.529"
## [1] "2023-12-10 00:26:15 MST"
## [1] "Linear Regression with Forward Selection Model has been fitted"
## [1] "Best R2: 0.53"
## [1] "2023-12-10 00:26:15 MST"
## [1] "GAM with Splines has been fitted"
## [1] "Best R2: 0.564"
## [1] "2023-12-10 00:28:18 MST"
## [1] "Random Forest has been fitted"
## [1] "Best R2: 0.634"
## [1] "2023-12-10 00:29:18 MST"
## [1] "Boosted Trees has been fitted"
## [1] "Best R2: 0.69"
## [1] "2023-12-10 00:29:46 MST"
## [1] "Artificial Neural Net has been fitted"
## [1] "Best R2: 0.542"
## [1] "----->Gradient Boosted Trees works best"
## [1] "Final RSME: 3.85142184682475e-05"
## [1] "Final R2: 0.682442559816678"
## [1] "Overall Best Model: Gradient Boosted Trees"
## [1] "2023-12-10 00:29:52 MST"

```

#Fit final model

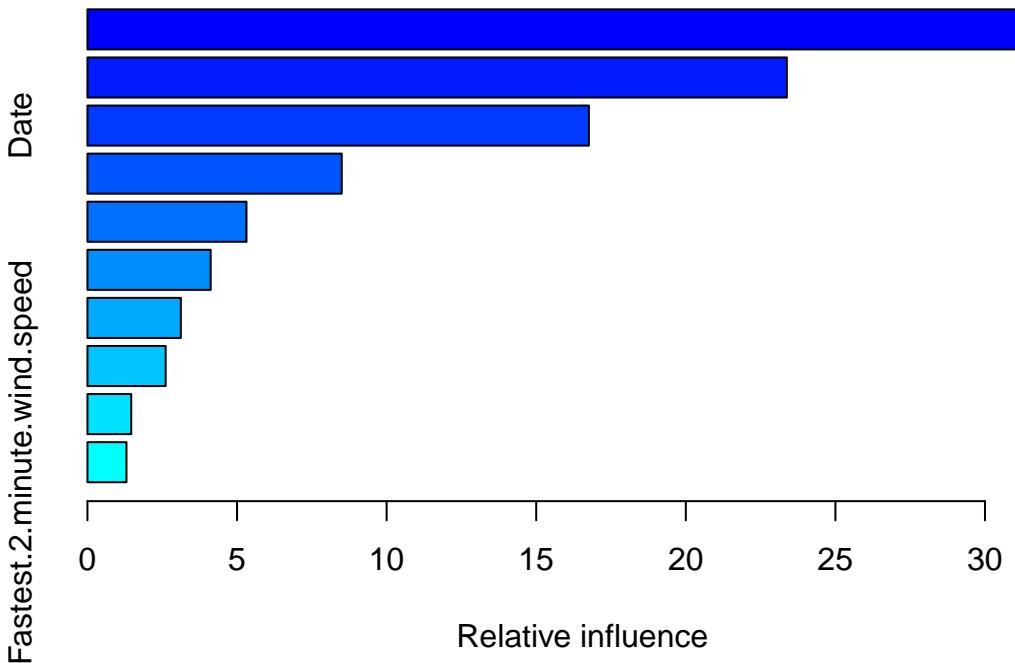
```

man_trans_final_pars = data.frame(man_trans_cv_results[3])
training = trainControl(method = "cv", number = 10)
man_trans_finalModelGrid <- data.frame(interaction.depth = man_trans_final_pars$interaction.depth,
                                         n.trees = man_trans_final_pars$n.trees,
                                         shrinkage = man_trans_final_pars$shrinkage,
                                         n.minobsinnode = man_trans_final_pars$n.minobsinnode) # use default

# cross-validation of Boosted Tree
man_trans_fit_final = train(MAN_Ozone_8.HR.RUN.AVG-BEGIN.HOUR_Arithmetic.Mean ~ .,
                            data = man_transformed,
                            verbose=FALSE,
                            method = "gbm",
                            trControl = training,
                            tuneGrid = man_trans_finalModelGrid)

summary(man_trans_fit_final)

```



```

## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean
## Date Date
## Average.Mixing.Height.power Average.Mixing.Height.power
## log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean
## log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean
## Direction.of.fastest.2.minute.wind Direction.of.fastest.2.minute.wind
## Fastest.2.minute.wind.speed Fastest.2.minute.wind.speed

## rel.inf
## HW24_Outdoor.Temperature_1.HOUR_Arithmetic.Mean 33.421554
## HW24_Relative.Humidity_1.HOUR_Arithmetic.Mean 23.376478
## Date 16.759888
## Average.Mixing.Height.power 8.500795
## log_HW24_SO2.max.5.min.avg_1.HOUR_Arithmetic.Mean 5.316116
## log_HW24_Carbon.monoxide_8.HR.RUN.AVG.END.HOUR_Arithmetic.Mean 4.119937
## HW24_Std.Dev.Hz.Wind.Direction_1.HOUR_Arithmetic.Mean 3.122461
## HW24_Wind.Direction...Resultant_1.HOUR_Arithmetic.Mean 2.614680
## Direction.of.fastest.2.minute.wind 1.463324
## Fastest.2.minute.wind.speed 1.304768

man_trans_fit_final

## Stochastic Gradient Boosting

```

```
##  
## 2017 samples  
##    10 predictor  
##  
## No pre-processing  
## Resampling: Cross-Validated (10 fold)  
## Summary of sample sizes: 1816, 1815, 1815, 1817, 1815, 1816, ...  
## Resampling results:  
##  
##    RMSE      Rsquared   MAE  
##    0.005338716  0.693069  0.004176643  
##  
## Tuning parameter 'n.trees' was held constant at a value of 300  
## Tuning  
##  
## Tuning parameter 'shrinkage' was held constant at a value of 0.1  
##  
## Tuning parameter 'n.minobsinnode' was held constant at a value of 10
```

```
#run all above
```