Ozone_Random_Forest

Ryan Erickson

2023-06-05

Libraries

```
library(dplyr)
library(tidyr)
library(MASS)
library(rpart)
library(ranger)
library(pROC)
library(Metrics)
library(RColorBrewer)
```

Data

```
ozone_data=read.csv("../final_data/ozone_data.csv")%>%
  dplyr::select(site_name,date,lat,long,mda8,everything())

ozone_data$site_name = as.factor(ozone_data$site_name)
ozone_data$date = as.factor(ozone_data$date)

ozone_data_no.mda8_names = ozone_data %>%
  separate(date, c("Months","Years"),remove =F) %>%
  mutate(date=as.character.POSIXt(date),
         Months=match(Months,month.abb)) %>%
  group_by(Months) %>%
  arrange(Years, Months)  %>%
  ungroup() %>%
  dplyr::select(-c("mda8","site_name","date","Months","Years","X")) %>%
  as.data.frame()

ozone_data_mda8_first.no_name = ozone_data %>%
  separate(date, c("Months","Years"),remove =F) %>%
  mutate(date=as.character.POSIXt(date),
         Months=match(Months,month.abb)) %>%
  group_by(Months) %>%
  arrange(Years, Months)  %>%
  ungroup() %>%
  dplyr::select(mda8,everything()) %>%
  dplyr::select(-c("site_name","date","Months","Years","X")) %>%
  as.data.frame()

ozone_data %>%
  separate(date, c("Months","Years"),remove =F) %>%
  mutate(date=as.character.POSIXt(date),
         Months=match(Months,month.abb)) %>%
  group_by(Months) %>%
  arrange(Years, Months)  %>%
  ungroup() %>%
  dplyr::select(mda8,everything()) %>%
  dplyr::select(-c("Months","Years","X")) %>%
  write.csv("../final_data/ozone_data_sorted.csv")

head(ozone_data_no.mda8_names)
```

```
##       lat     long       ndvi    elev dist2road road_length     tmax    rhmax
## 1 4387727 536954.6 0.11474641 1793.14  11202.39       0.000 289.8151 64.13946
## 2 4435552 481219.8 0.13446639 1593.88    544.80       0.000 290.4293 66.00546
## 3 4400142 501060.2 0.07355672 1609.04   1174.89       0.000 291.2932 62.66088
## 4 4379800 503676.9 0.08371748 1746.35    280.62    3251.268 290.1372 62.51497
## 5 4403283 499556.4 0.17963080 1609.04    413.00    1231.818 291.2932 62.66088
## 6 4399329 484750.3 0.13003676 1766.56   1430.77       0.000 288.9421 63.72078
##        pmax apr_dummy may_dummy jun_dummy jul_dummy aug_dummy sep_dummy
## 1 0.7027266         1         0         0         0         0         0
## 2 0.8995762         1         0         0         0         0         0
## 3 0.9036559         1         0         0         0         0         0
## 4 1.0217647         1         0         0         0         0         0
## 5 0.9036559         1         0         0         0         0         0
## 6 0.9148828         1         0         0         0         0         0
##   oct_dummy yr_2018_dummy yr_2019_dummy yr_2020_dummy yr_2021_dummy
## 1         0             1             0             0             0
## 2         0             1             0             0             0
## 3         0             1             0             0             0
## 4         0             1             0             0             0
## 5         0             1             0             0             0
## 6         0             1             0             0             0
##   yr_2022_dummy
## 1             0
## 2             0
## 3             0
## 4             0
## 5             0
## 6             0
```

```
head(ozone_data_mda8_first.no_name)
```

```
##       mda8     lat     long       ndvi    elev dist2road road_length     tmax
## 1 45.43727 4387727 536954.6 0.11474641 1793.14  11202.39       0.000 289.8151
## 2 43.20485 4435552 481219.8 0.13446639 1593.88    544.80       0.000 290.4293
## 3 37.63830 4400142 501060.2 0.07355672 1609.04   1174.89       0.000 291.2932
## 4 45.89848 4379800 503676.9 0.08371748 1746.35    280.62    3251.268 290.1372
## 5 37.24626 4403283 499556.4 0.17963080 1609.04    413.00    1231.818 291.2932
## 6 43.74330 4399329 484750.3 0.13003676 1766.56   1430.77       0.000 288.9421
##      rhmax      pmax apr_dummy may_dummy jun_dummy jul_dummy aug_dummy
## 1 64.13946 0.7027266         1         0         0         0         0
## 2 66.00546 0.8995762         1         0         0         0         0
## 3 62.66088 0.9036559         1         0         0         0         0
## 4 62.51497 1.0217647         1         0         0         0         0
## 5 62.66088 0.9036559         1         0         0         0         0
## 6 63.72078 0.9148828         1         0         0         0         0
##   sep_dummy oct_dummy yr_2018_dummy yr_2019_dummy yr_2020_dummy yr_2021_dummy
## 1         0         0             1             0             0             0
## 2         0         0             1             0             0             0
## 3         0         0             1             0             0             0
## 4         0         0             1             0             0             0
## 5         0         0             1             0             0             0
## 6         0         0             1             0             0             0
##   yr_2022_dummy
## 1             0
## 2             0
## 3             0
## 4             0
## 5             0
## 6             0
```

Summary of LM models with Dummy Variables Included

```
line1 = glm(mda8~., data=ozone_data_mda8_first.no_name)
summary(line1)
```

```
##
## Call:
## glm(formula = mda8 ~ ., data = ozone_data_mda8_first.no_name)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -6.604  -1.639   0.006   1.554   7.727
##
## Coefficients: (2 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -5.240e+02  1.993e+02  -2.629 0.009093 **
## lat            9.601e-05  3.423e-05   2.805 0.005422 **
## long          -1.766e-04  5.990e-05  -2.948 0.003495 **
## ndvi          -9.684e+00  3.729e+00  -2.597 0.009960 **
## elev           5.089e-02  8.326e-03   6.113 3.71e-09 ***
## dist2road      6.616e-04  3.088e-04   2.143 0.033084 *
## road_length    1.132e-03  2.846e-04   3.978 9.10e-05 ***
## tmax           4.710e-01  1.442e-01   3.266 0.001244 **
## rhmax         -4.193e-02  4.263e-02  -0.983 0.326331
## pmax          -8.744e-01  3.226e-01  -2.711 0.007179 **
## apr_dummy      1.167e+01  6.210e-01  18.787  < 2e-16 ***
## may_dummy      1.103e+01  1.130e+00   9.769  < 2e-16 ***
## jun_dummy      1.087e+01  1.872e+00   5.809 1.89e-08 ***
## jul_dummy      1.332e+01  2.355e+00   5.657 4.17e-08 ***
## aug_dummy      1.361e+01  2.144e+00   6.349 1.00e-09 ***
## sep_dummy      5.509e+00  1.554e+00   3.545 0.000468 ***
## oct_dummy             NA         NA      NA       NA
## yr_2018_dummy -9.231e-01  5.206e-01  -1.773 0.077409 .
## yr_2019_dummy -1.307e+00  5.433e-01  -2.405 0.016887 *
## yr_2020_dummy -1.197e+00  5.188e-01  -2.308 0.021810 *
## yr_2021_dummy  2.748e+00  5.329e-01   5.157 5.09e-07 ***
## yr_2022_dummy         NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 6.536941)
##
##     Null deviance: 17900.9  on 271  degrees of freedom
## Residual deviance:  1647.3  on 252  degrees of freedom
## AIC: 1303.8
##
## Number of Fisher Scoring iterations: 2
```

Summary of LM models with no Dummy Variables

```
line_nd=ozone_data_mda8_first.no_name[,c(-2,-3,-11:-22)]
line = glm(mda8~., data=line_nd)
summary(line)
```

```
##
## Call:
## glm(formula = mda8 ~ ., data = line_nd)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.9213   -2.9066    0.0289    3.2092   13.8076
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.275e+02  2.085e+01 -15.711  < 2e-16 ***
## ndvi         1.173e+01  6.073e+00   1.931   0.0545 .
## elev         6.946e-02  5.891e-03  11.791  < 2e-16 ***
## dist2road   -6.172e-04  1.341e-04  -4.602 6.51e-06 ***
## road_length -4.304e-04  3.398e-04  -1.267   0.2064
## tmax         8.320e-01  5.038e-02  16.515  < 2e-16 ***
## rhmax        6.522e-02  6.497e-02   1.004   0.3164
## pmax         9.255e-01  5.067e-01   1.826   0.0689 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 25.0794)
##
##     Null deviance: 17901  on 271  degrees of freedom
## Residual deviance:  6621  on 264  degrees of freedom
## AIC: 1658.2
##
```

```
## Number of Fisher Scoring iterations: 2
```

```
n = 21
qual_col_pals = brewer.pal.info[brewer.pal.info$category == 'qual',]
col_vector = unlist(mapply(brewer.pal, qual_col_pals$maxcolors, rownames(qual_col_pals)))
col=sample(col_vector, n)

for (i in 1:ncol(ozone_data_no.mda8_names)) {
 plot(x=ozone_data_no.mda8_names[,i],
       y=ozone_data_mda8_first.no_name$mda8,
       main = paste0('Figure ',i,': MDA8 vs. ',colnames(ozone_data_no.mda8_names)[i]),
       xlab = paste0(colnames(ozone_data_no.mda8_names)[i]),
       ylab = "MDA8 Value",
       pch = 19)
  abline(lm(reformulate(paste0(names(ozone_data_no.mda8_names[i])),"mda8"),ozone_data_mda8_first.no_nam
         col = col[i],
         lwd = 2)
}
```

Figure 1: MDA8 vs. lat

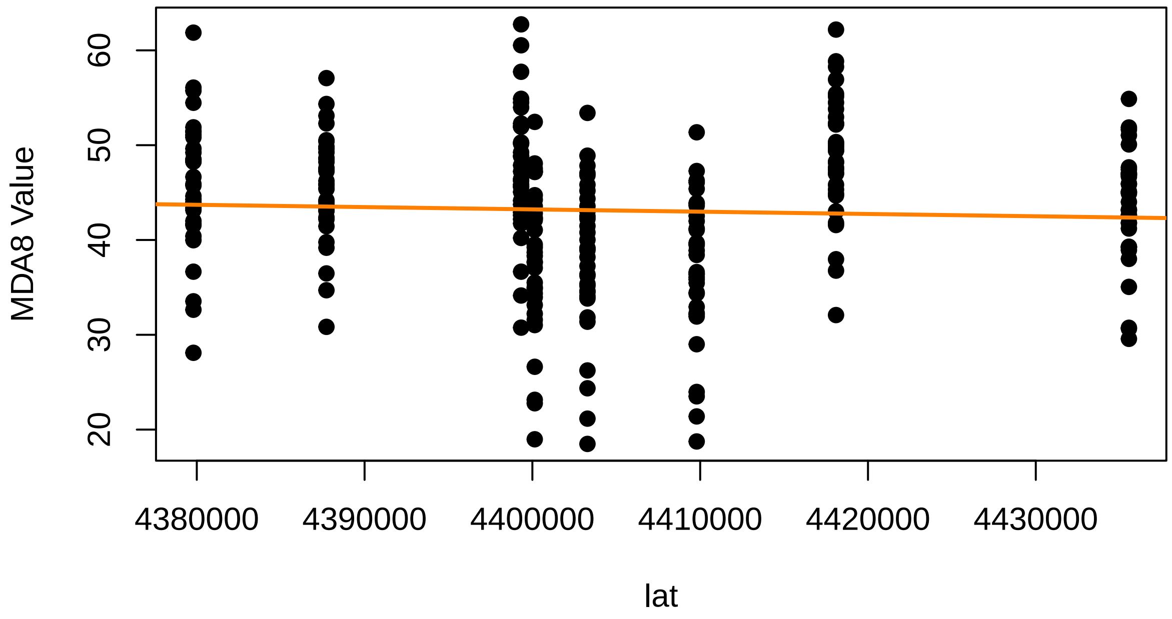


Figure 2: MDA8 vs. long

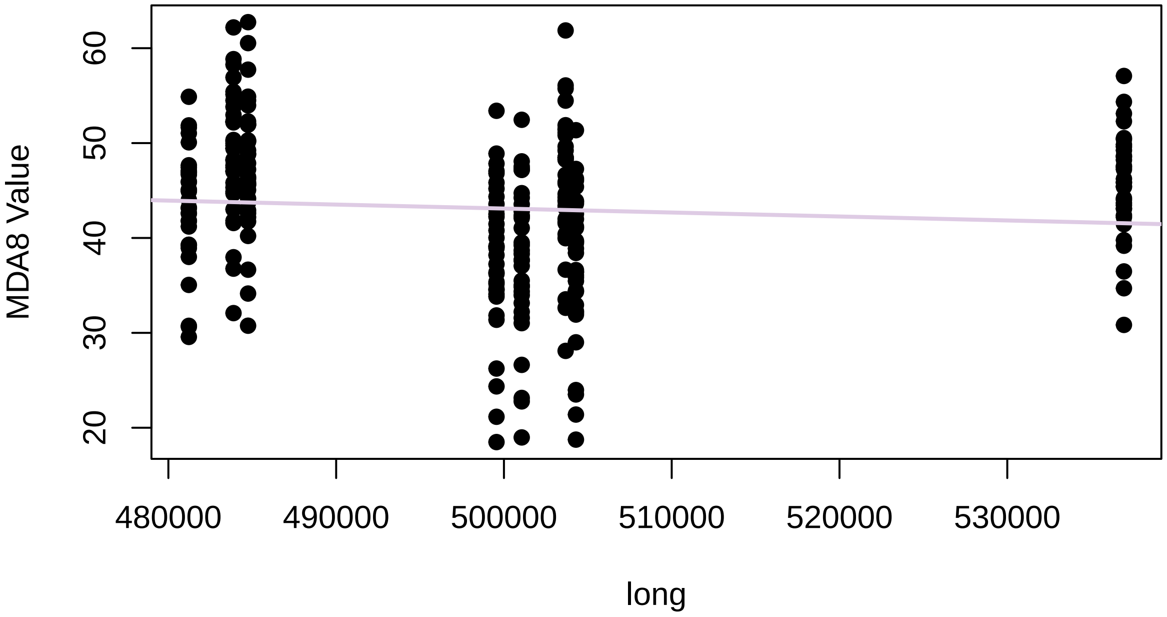


Figure 3: MDA8 vs. ndvi

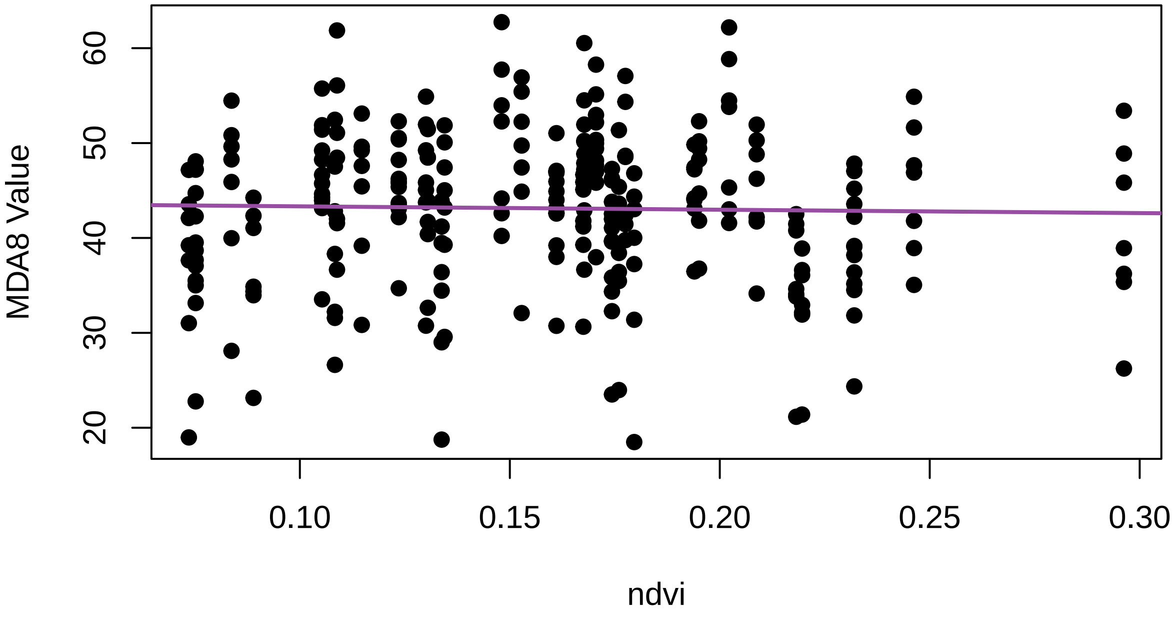


Figure 4: MDA8 vs. elev

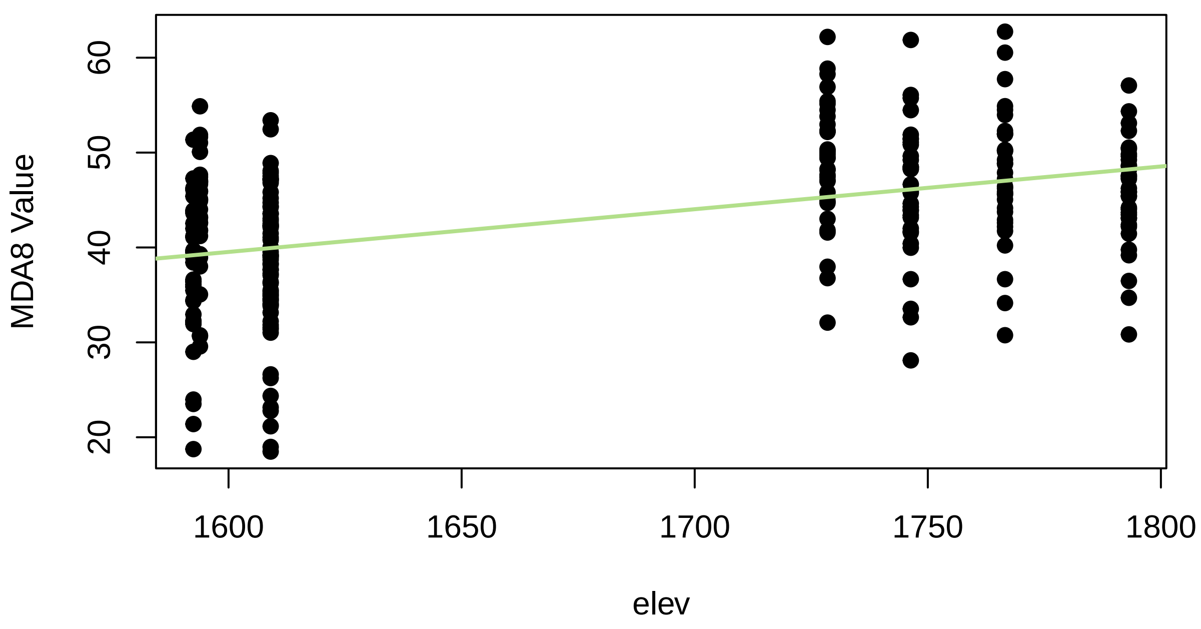


Figure 5: MDA8 vs. dist2road

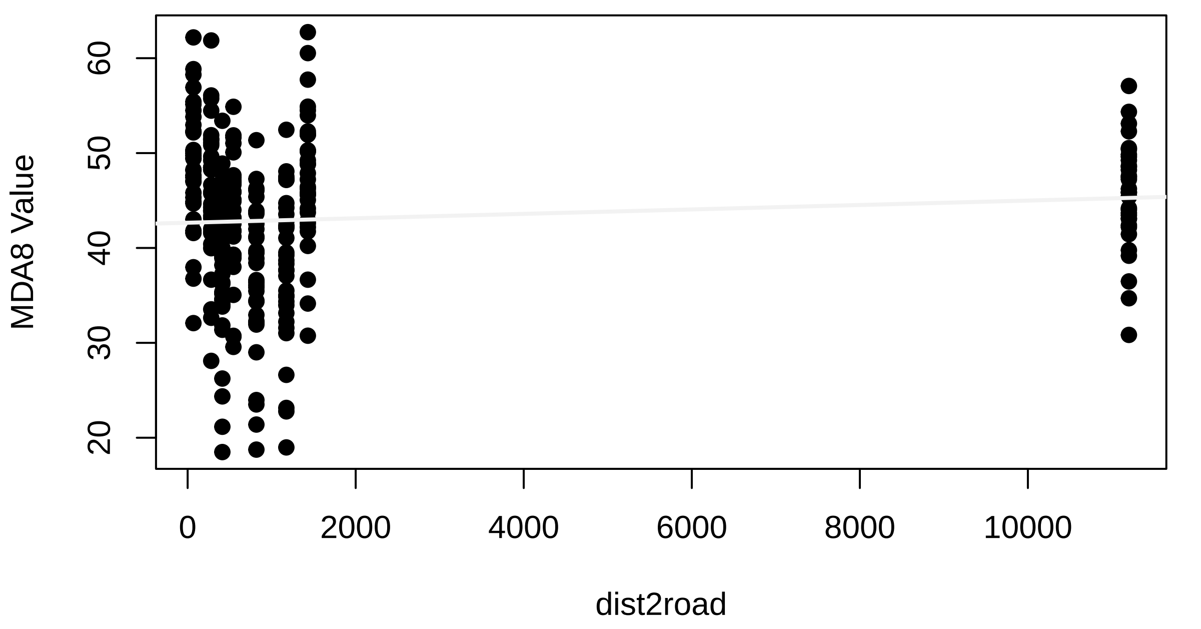


Figure 6: MDA8 vs. road_length

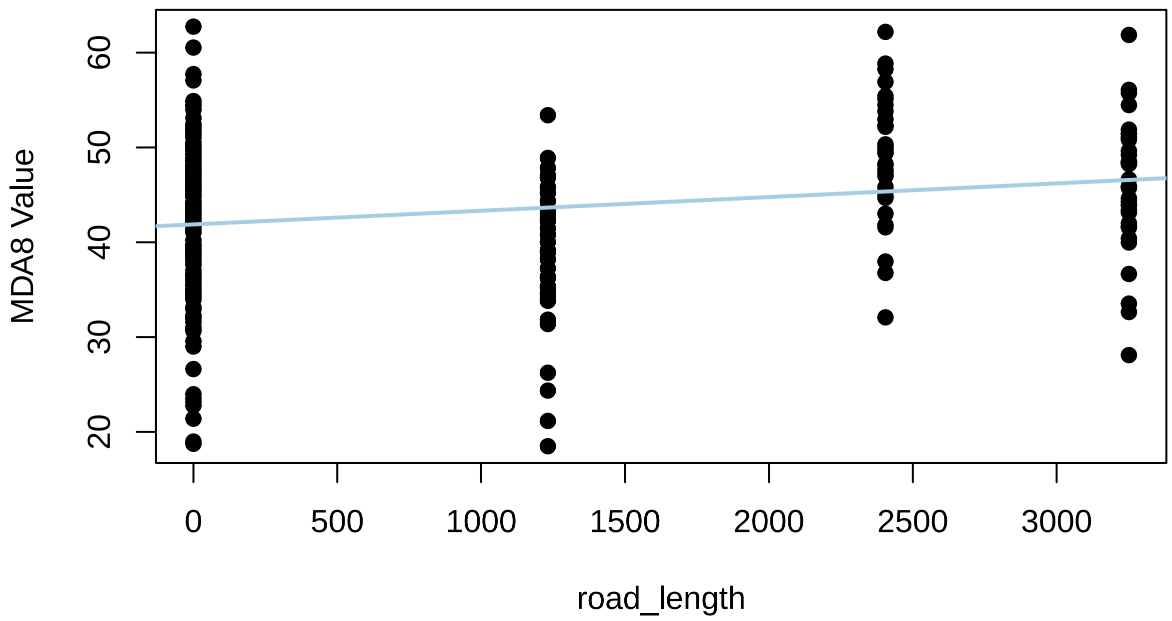


Figure 7: MDA8 vs. tmax

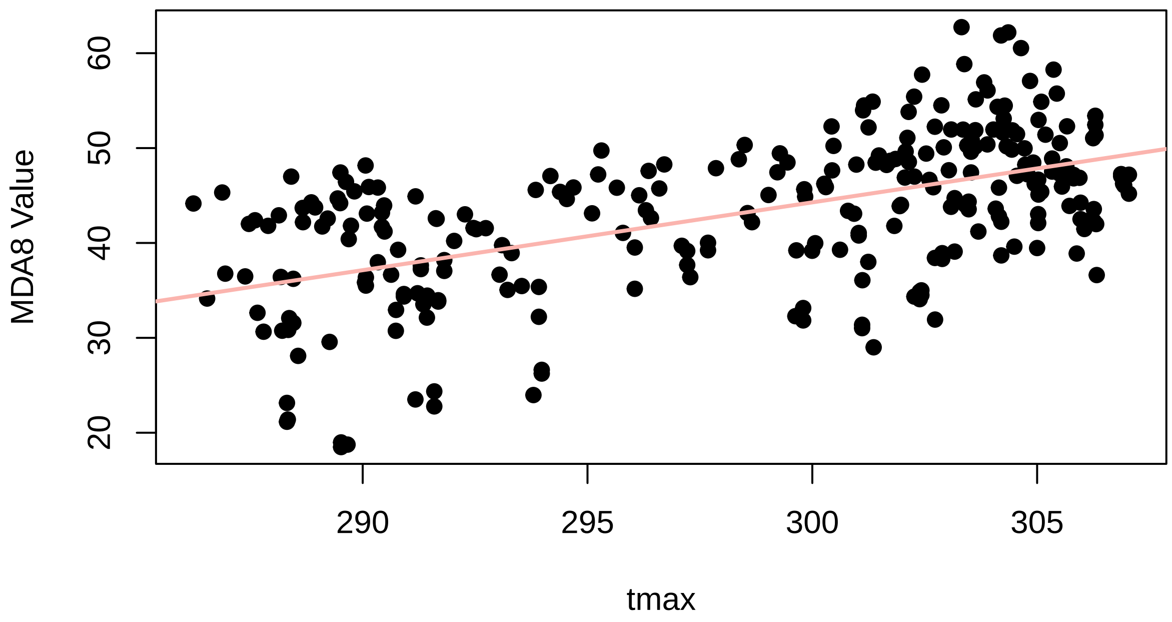


Figure 8: MDA8 vs. rhmax

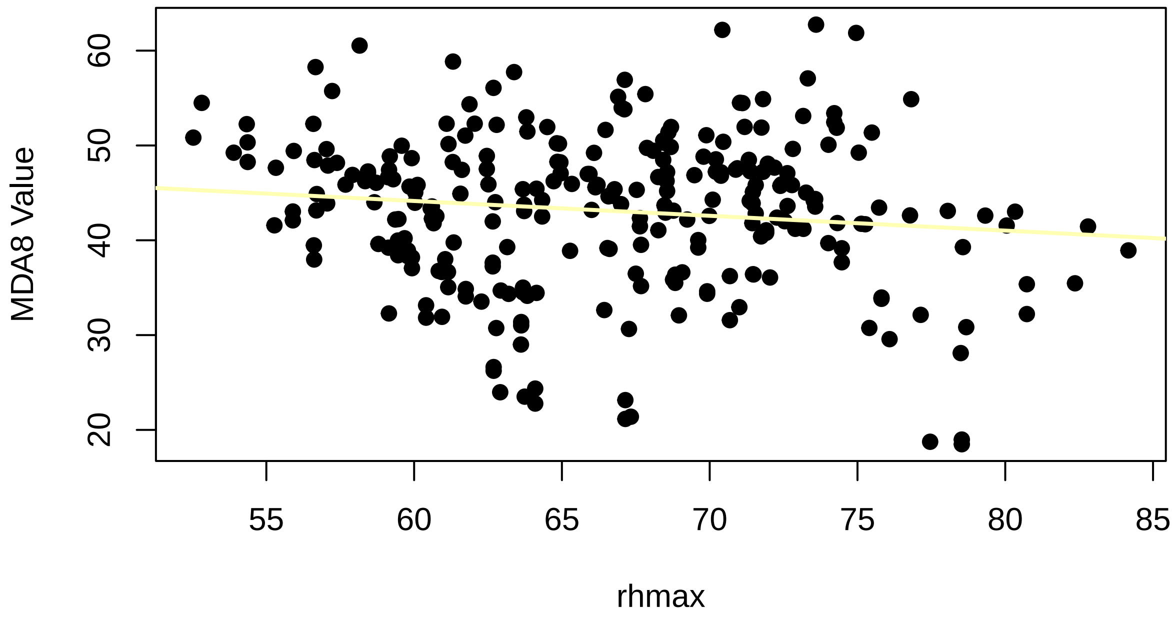


Figure 9: MDA8 vs. pmax

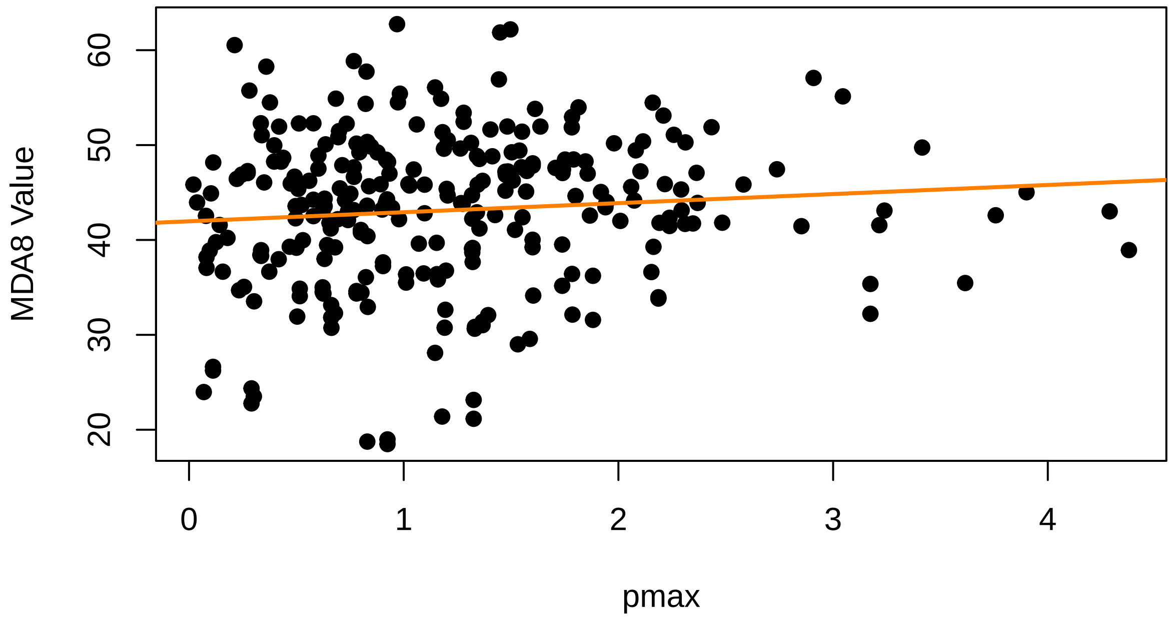


Figure 10: MDA8 vs. apr_dummy

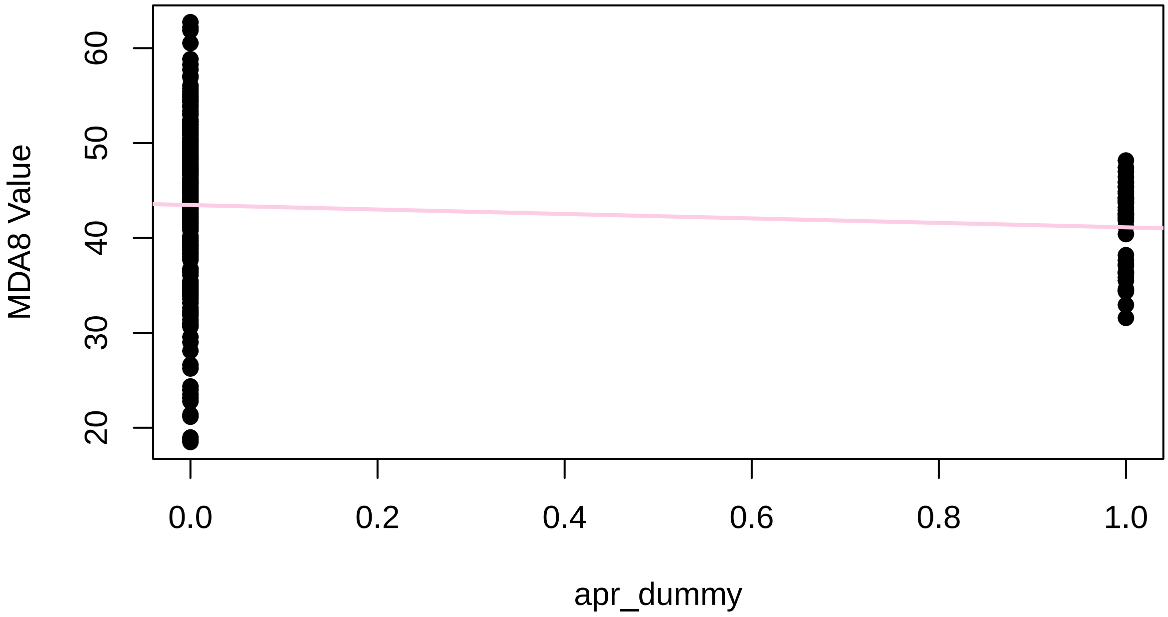


Figure 11: MDA8 vs. may_dummy

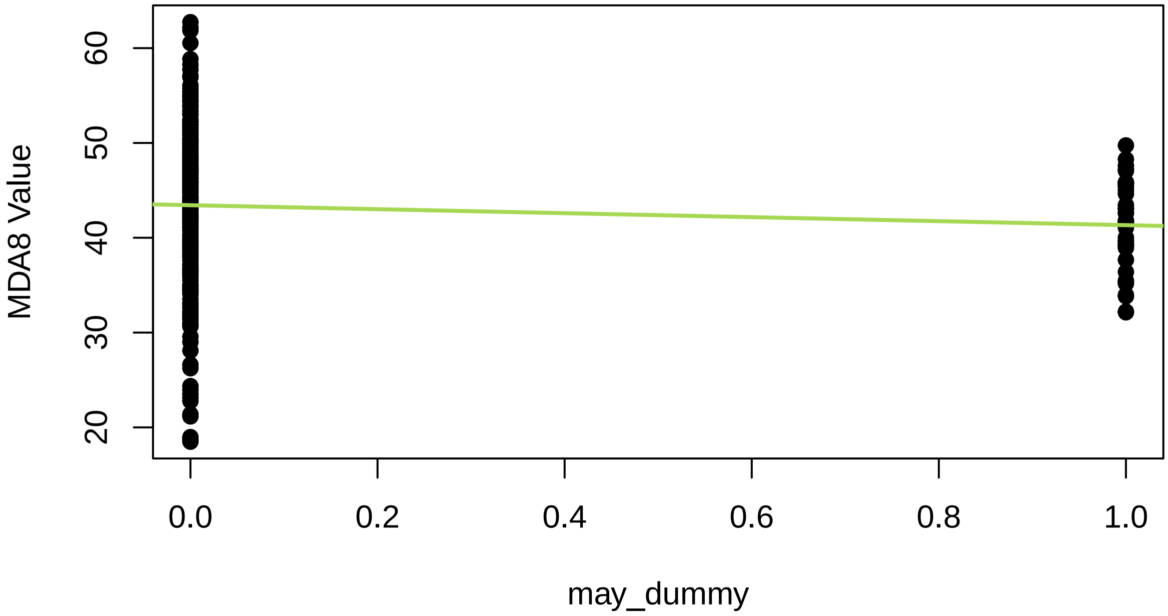


Figure 12: MDA8 vs. jun_dummy

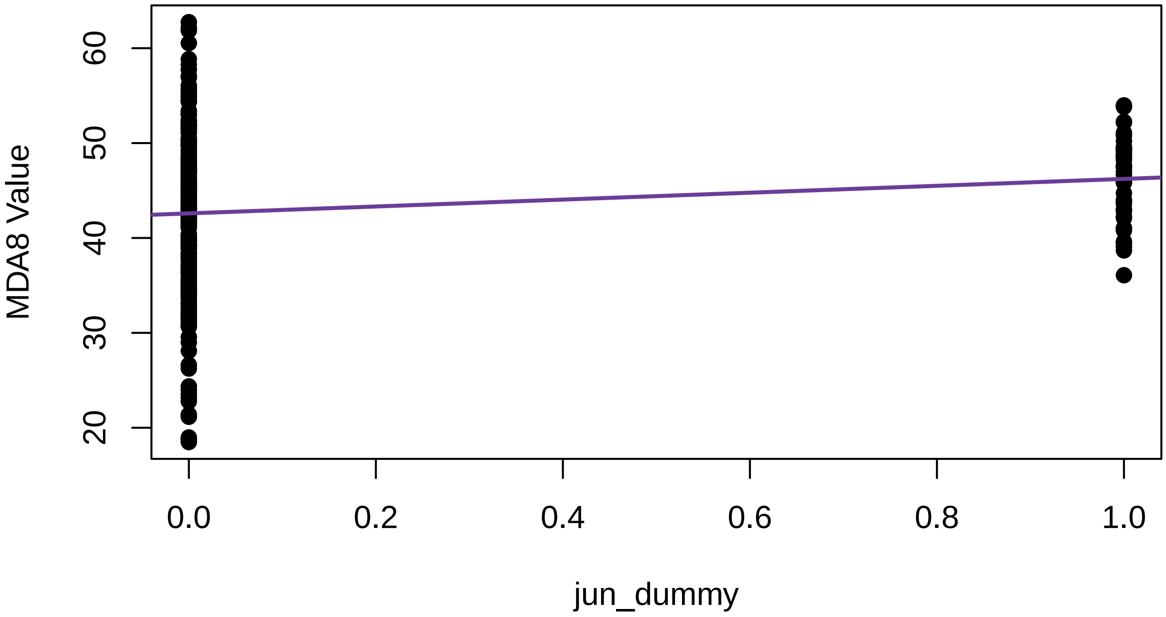


Figure 13: MDA8 vs. jul_dummy

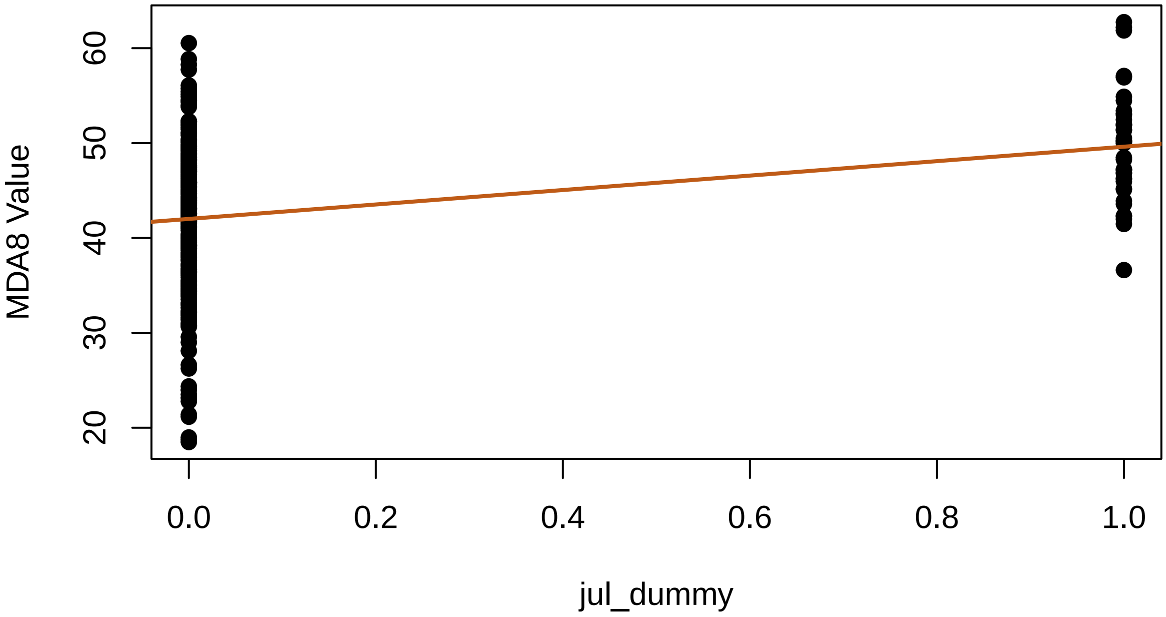


Figure 14: MDA8 vs. aug_dummy

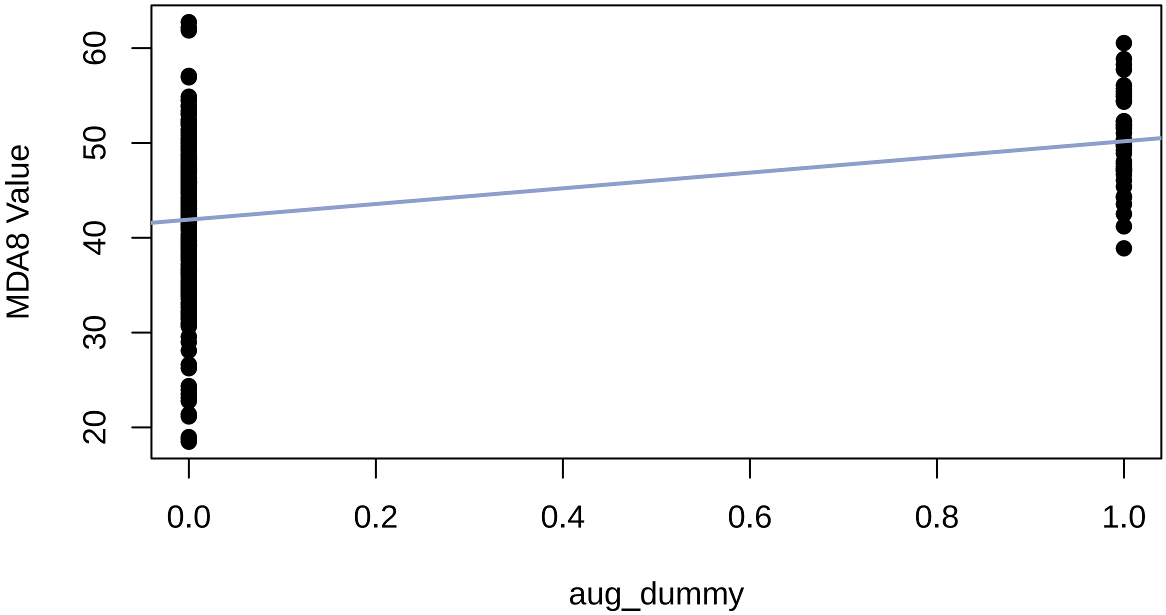


Figure 15: MDA8 vs. sep_dummy

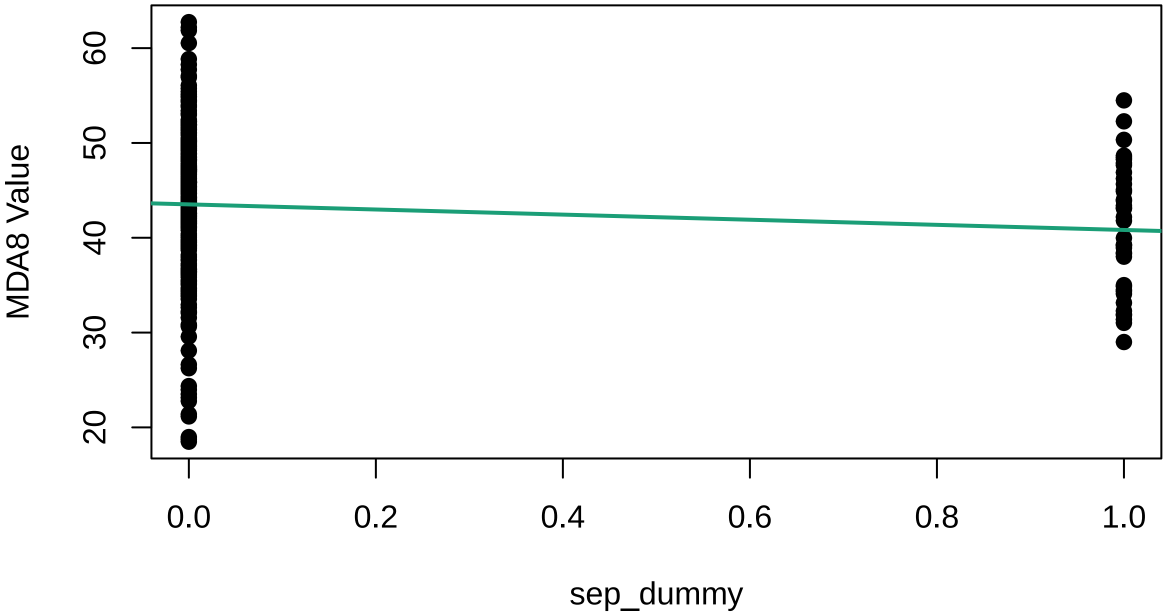


Figure 16: MDA8 vs. oct_dummy

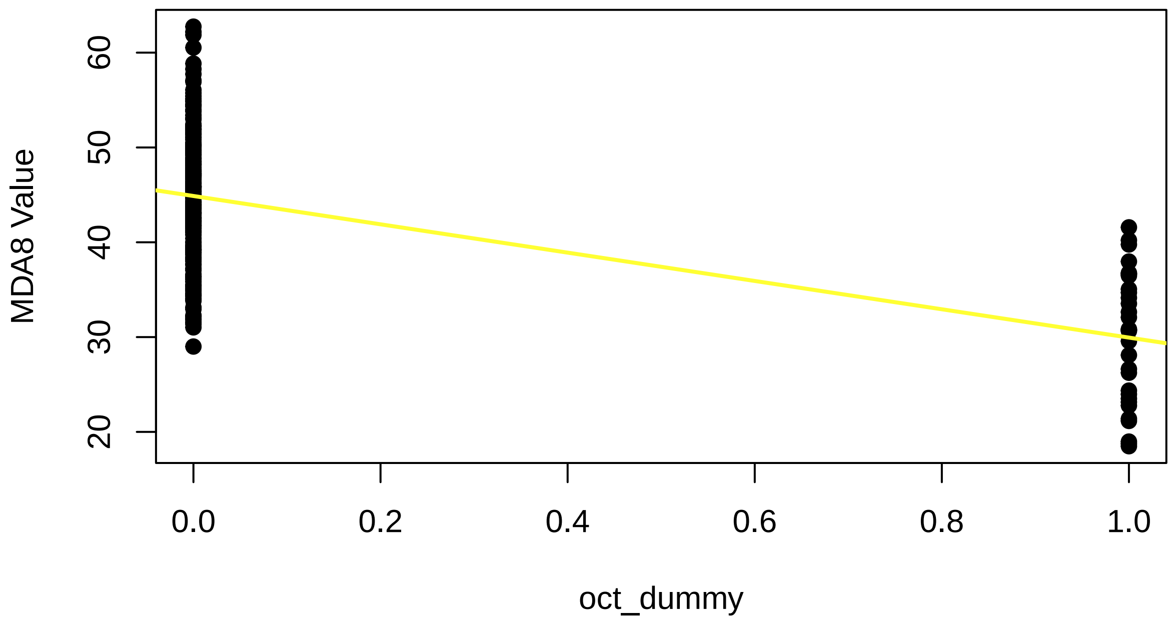


Figure 17: MDA8 vs. yr_2018_dummy

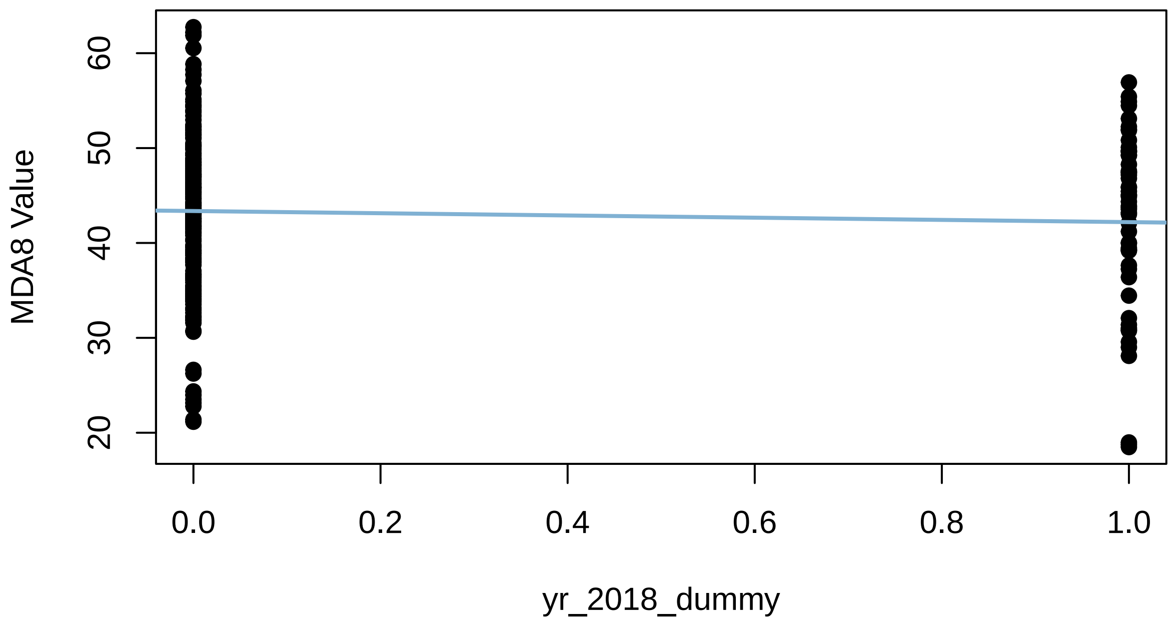


Figure 18: MDA8 vs. yr_2019_dummy

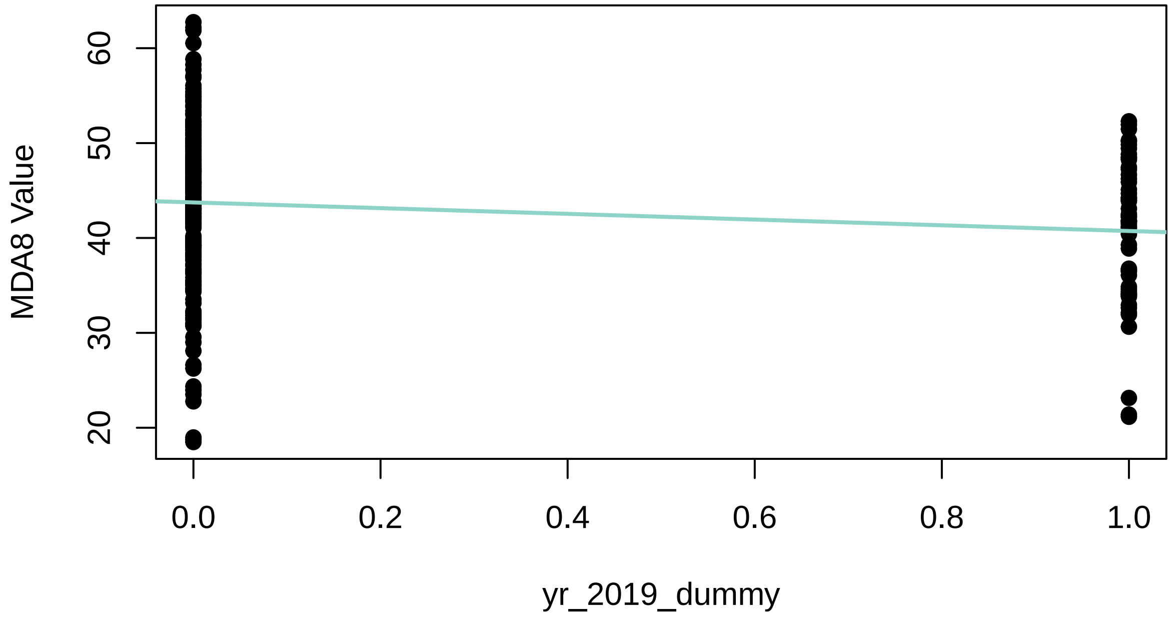


Figure 19: MDA8 vs. yr_2020_dummy

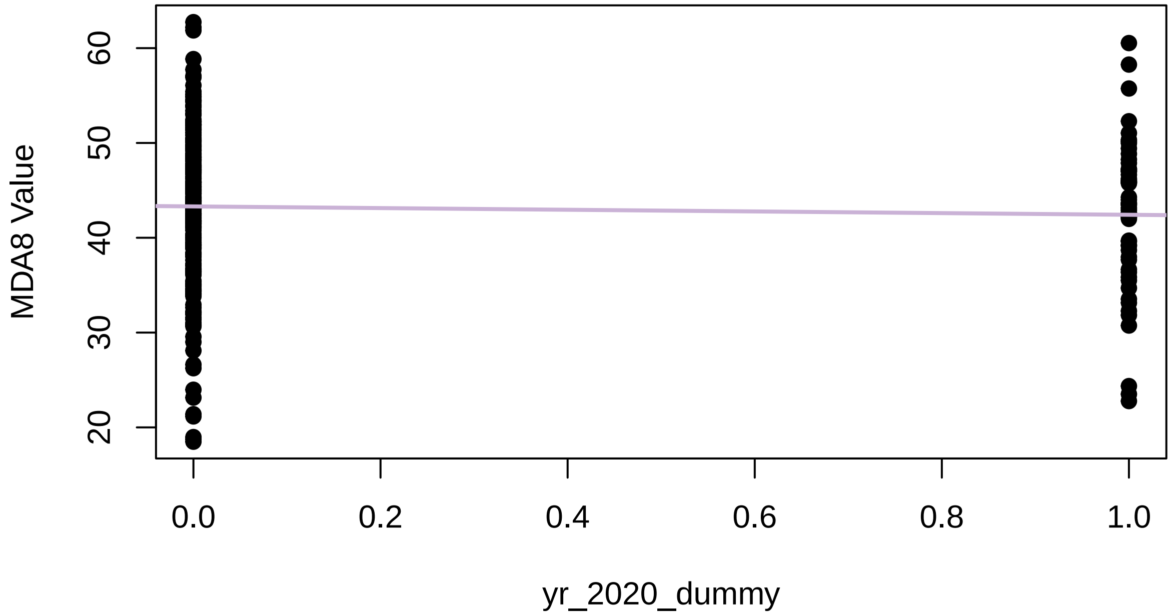


Figure 20: MDA8 vs. yr_2021_dummy

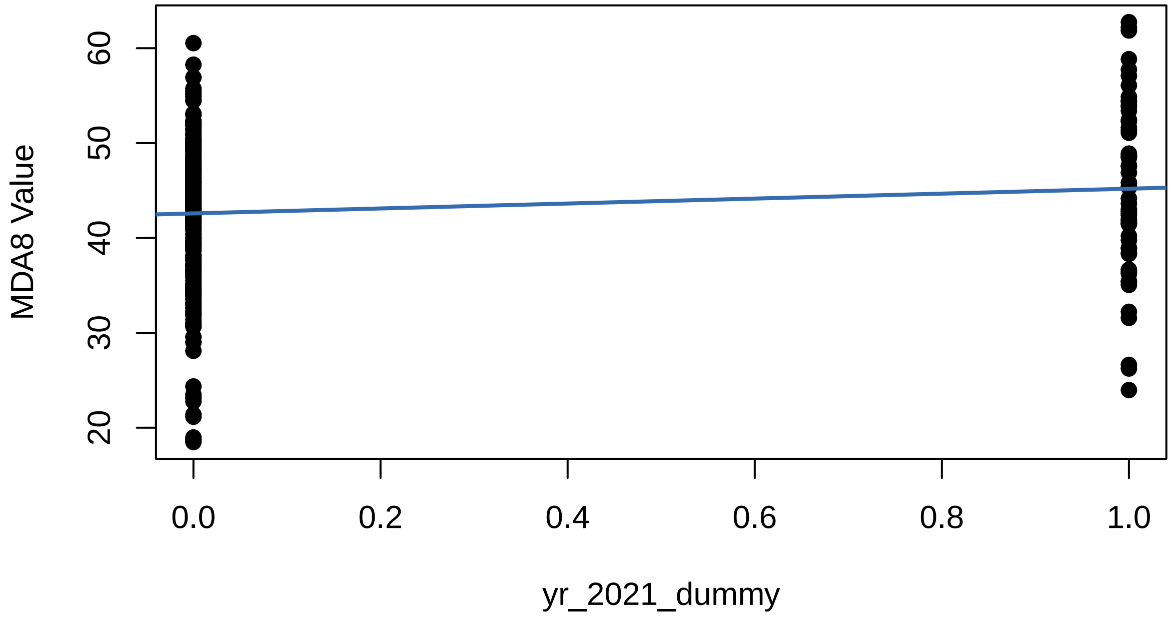
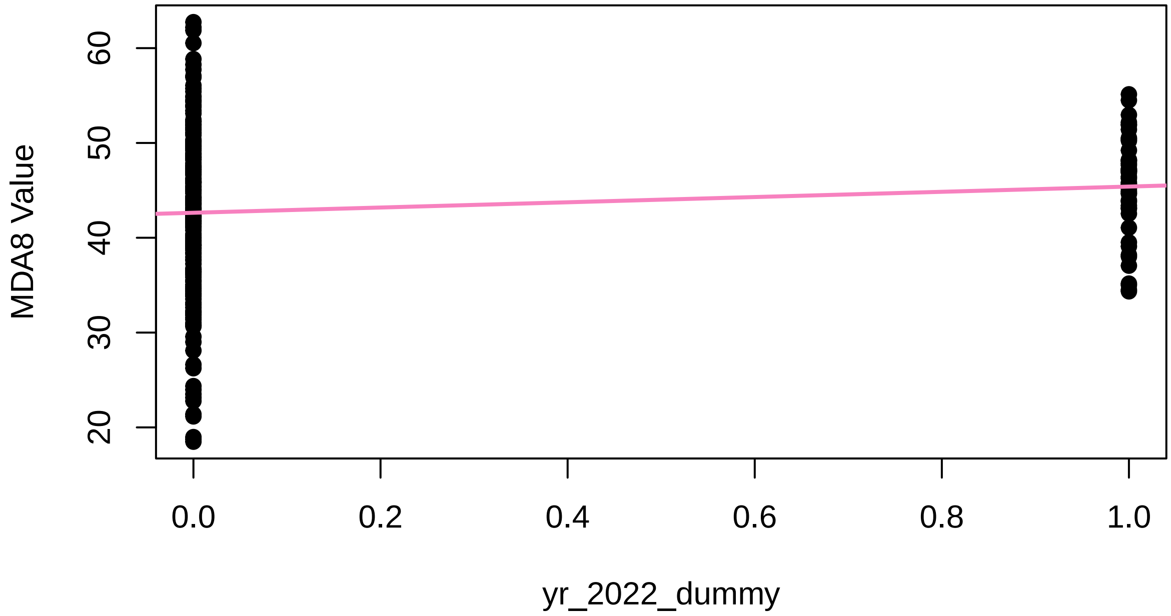


Figure 21: MDA8 vs. yr_2022_dummy



```
test=as.data.frame(ozone_data_mda8_first.no_name[,c(-2,-3,-11:-22)])
test1=as.data.frame(ozone_data_mda8_first.no_name[,c(-1,-2,-3,-11:-22)])
```

```
n = 7
qual_col_pals = brewer.pal.info[brewer.pal.info$category == 'qual',]
col_vector = unlist(mapply(brewer.pal, qual_col_pals$maxcolors, rownames(qual_col_pals)))
col=sample(col_vector, n)

for (i in 1:ncol(test1)) {
 plot(x=test1[,i],
       y=test$mda8,
       main = paste0('Figure ',i,': MDA8 vs. ',colnames(test1)[i]),
       xlab = paste0(colnames(test1)[i]),
       ylab = "MDA8 Value",
       pch = 19)
  abline(lm(reformulate(paste0(names(test1[i])),"mda8"),test),
         col = col[i],
         lwd = 2)
}
```

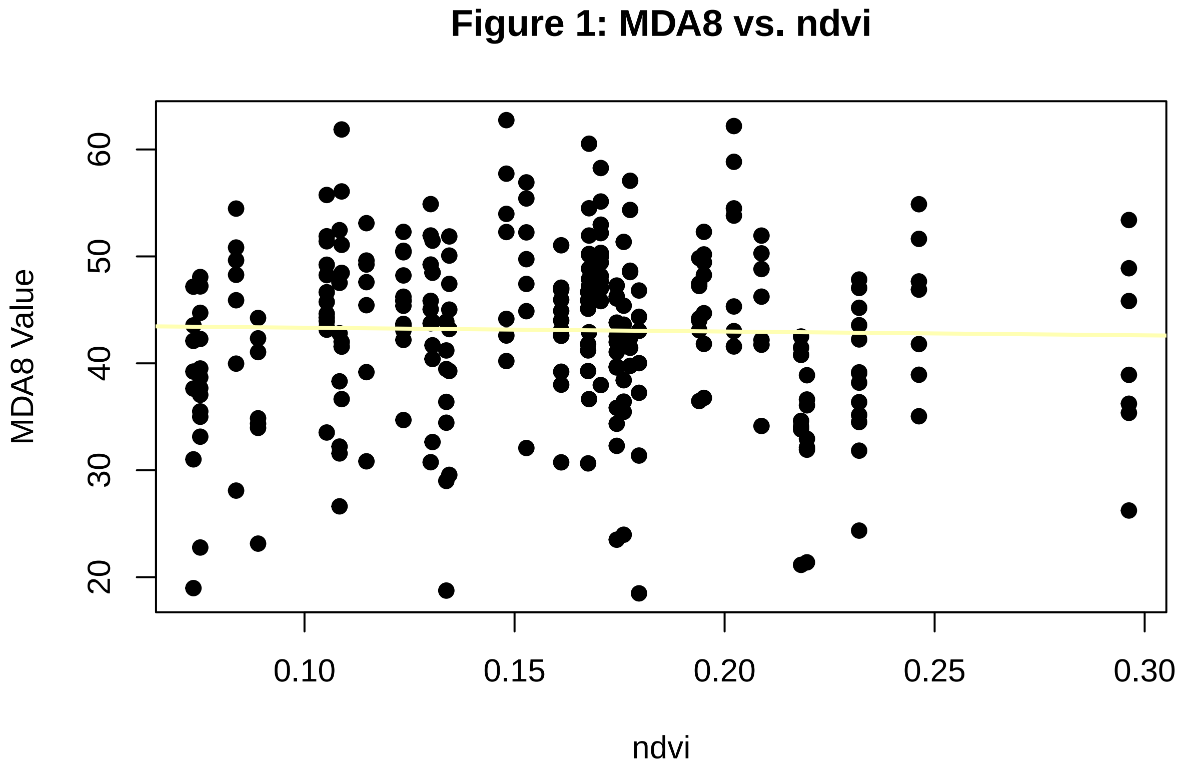


Figure 2: MDA8 vs. elev

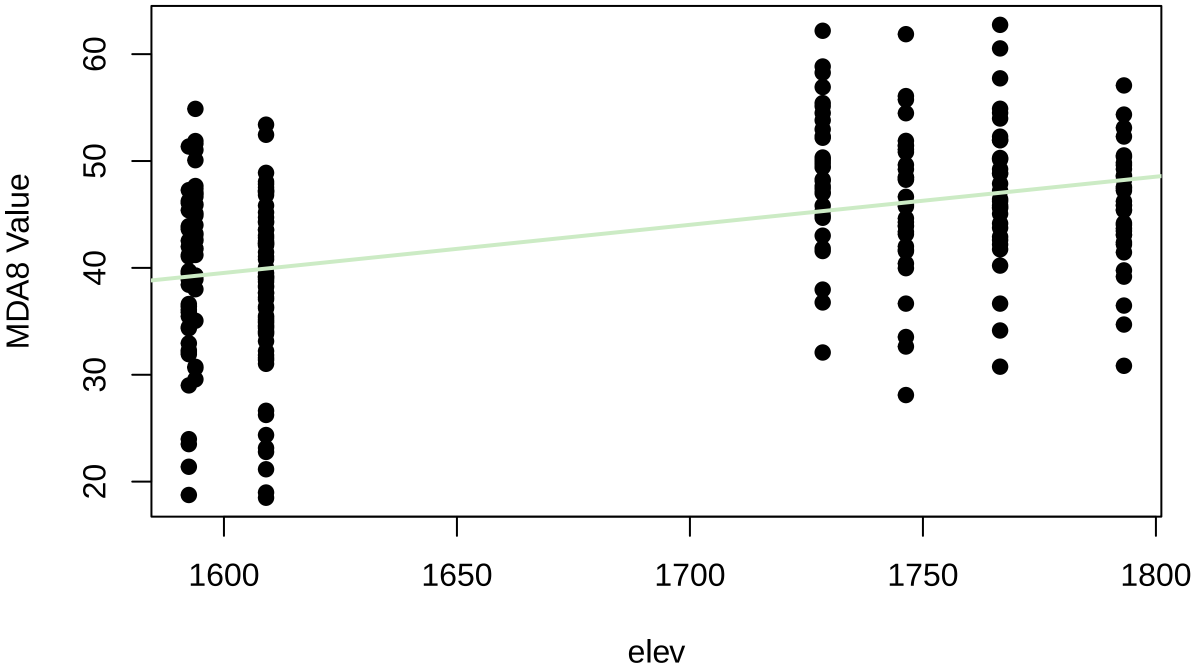
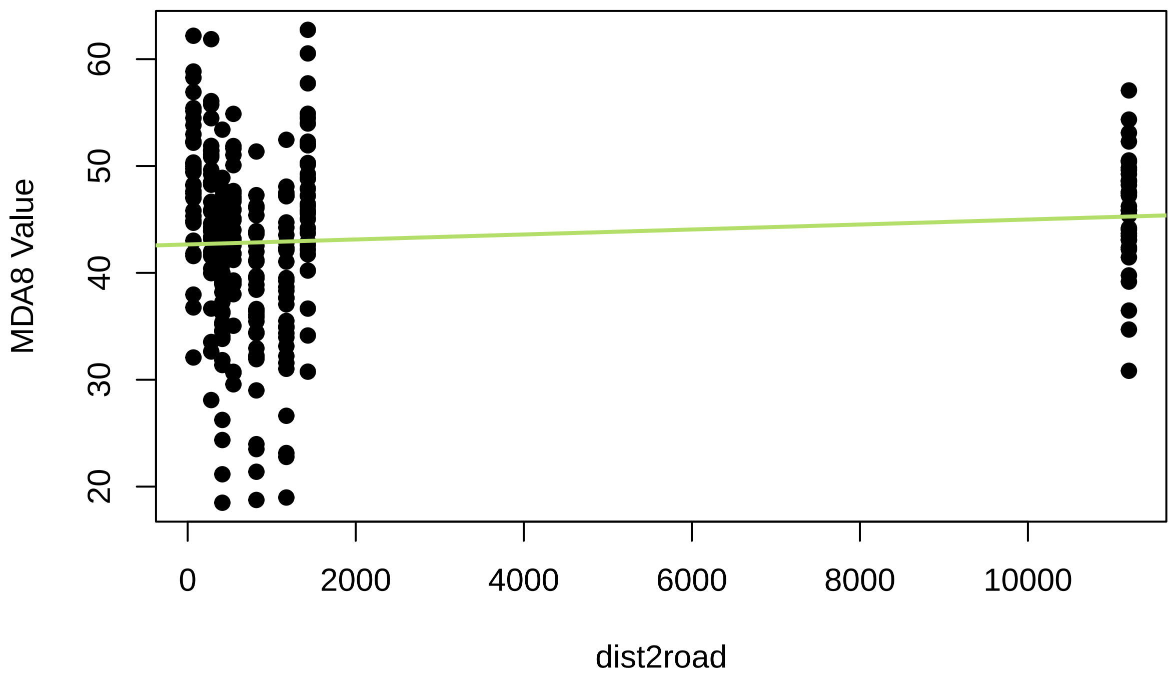


Figure 3: MDA8 vs. dist2road



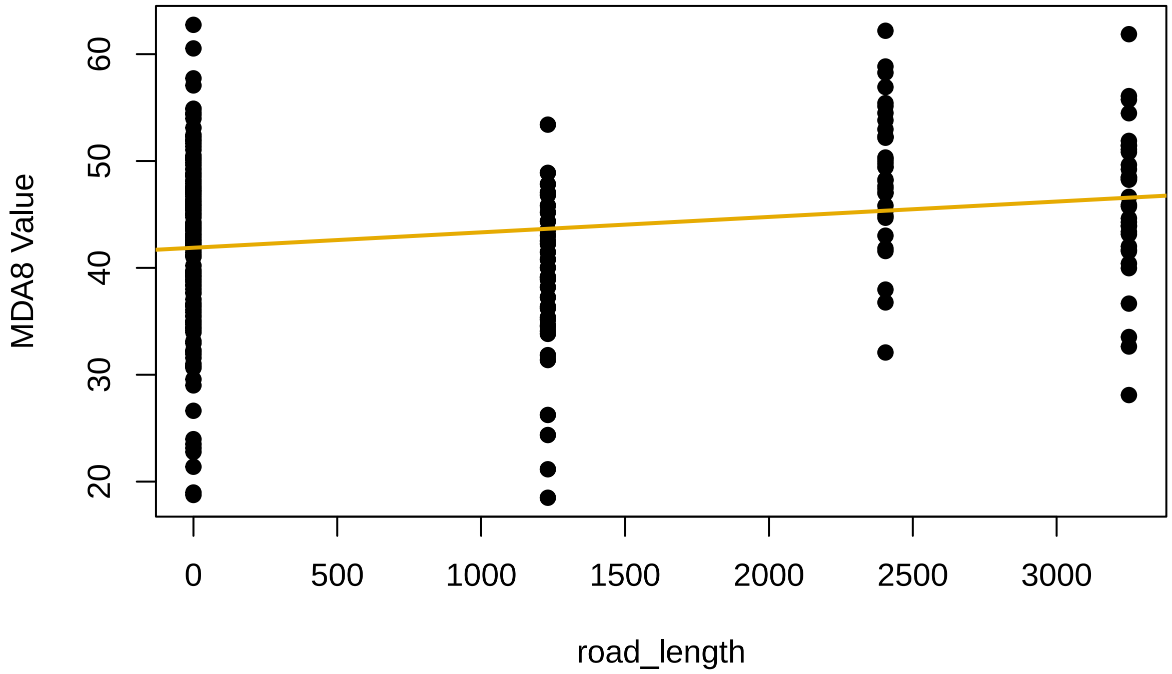
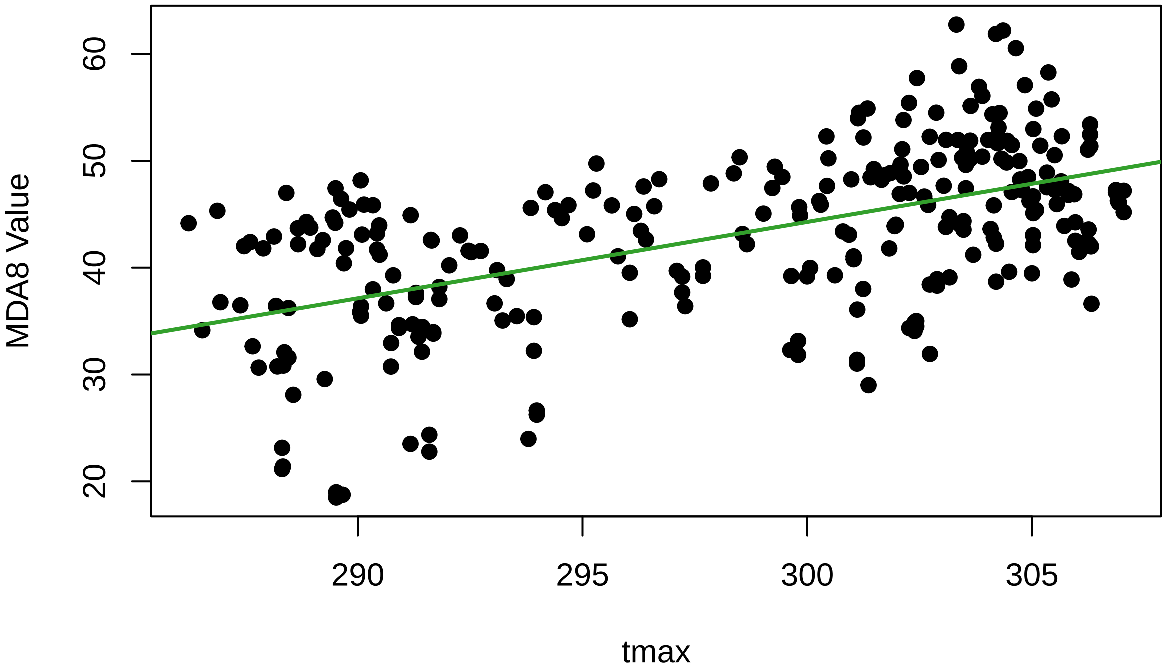


Figure 4: MDA8 vs. road_length

Figure 5: MDA8 vs. tmax



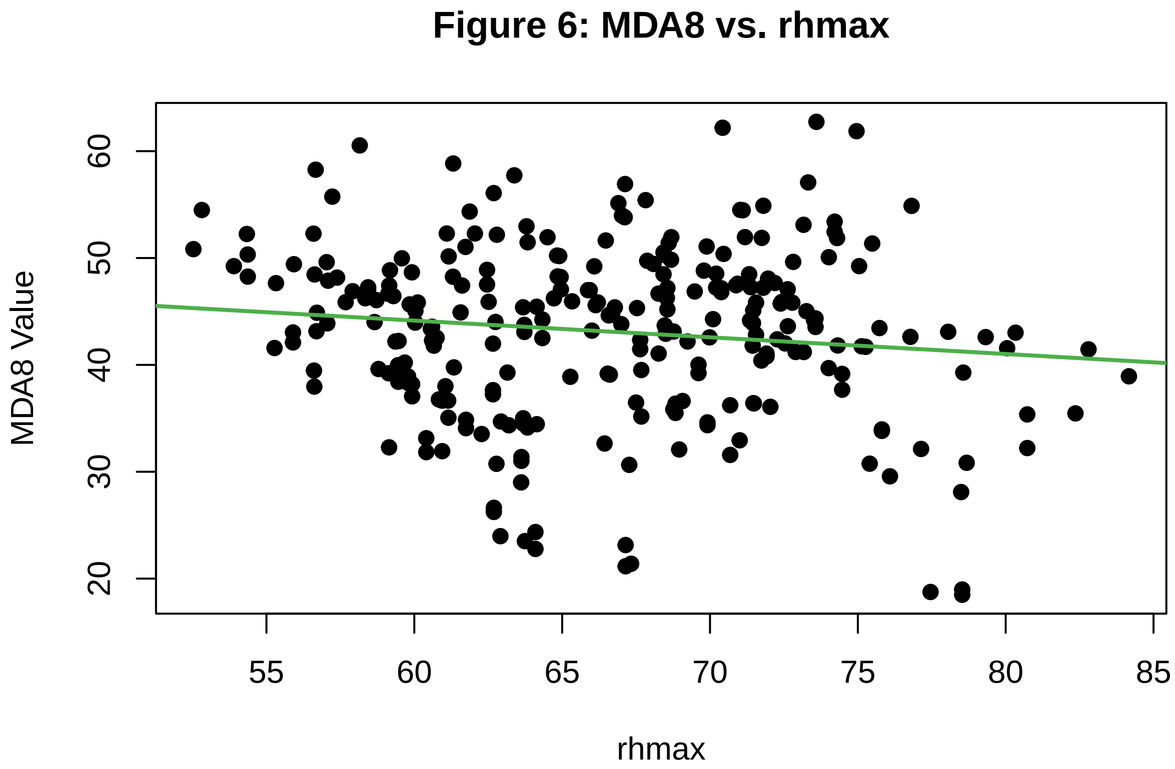
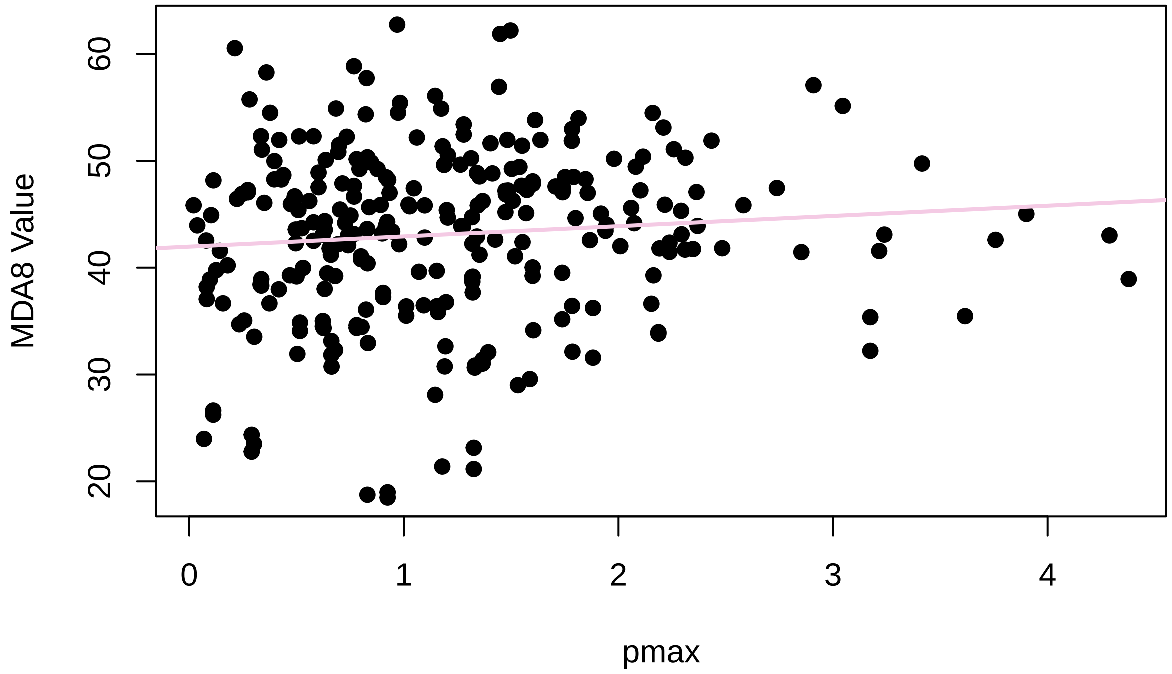


Figure 6: MDA8 vs. rhmax
MDA8 Value
20
30
40
50
60
55
60
65
70
75
80
85
rhmax

Figure 7: MDA8 vs. pmax



Splitting the Data - Testing Linear Models: 75% train/test split

```
sample_size = floor(0.75 * nrow(ozone_data_mda8_first.no_name))
set.seed(09111997)
split_dat = sample(seq_len(nrow(ozone_data_mda8_first.no_name)), size = sample_size, replace=FALSE)

ozone_train = ozone_data_mda8_first.no_name[split_dat, ]
ozone_test = ozone_data_mda8_first.no_name[-split_dat, ]

lm.final = glm(mda8~., data = ozone_train)
###predicting on training date with test data
pred.vals = predict(object=lm.final,new_data=ozone_test,type = "response")
summary(lm.final)
```

```
##
## Call:
## glm(formula = mda8 ~ ., data = ozone_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.2790  -1.5099  -0.0308   1.5132   7.0558
##
## Coefficients: (2 not defined because of singularities)
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -4.988e+02  2.373e+02  -2.102 0.036884 *
## lat            1.018e-04  4.090e-05   2.490 0.013666 *
## long          -1.891e-04  7.216e-05  -2.621 0.009505 **
## ndvi          -1.318e+01  4.490e+00  -2.935 0.003762 **
## elev           4.821e-02  9.898e-03   4.870 2.39e-06 ***
## dist2road      7.573e-04  3.716e-04   2.038 0.043010 *
## road_length    1.288e-03  3.390e-04   3.799 0.000197 ***
## tmax           3.337e-01  1.642e-01   2.032 0.043584 *
## rhmax         -3.335e-02  4.979e-02  -0.670 0.503918
## pmax          -1.128e+00  3.774e-01  -2.990 0.003168 **
## apr_dummy      1.127e+01  7.340e-01  15.353  < 2e-16 ***
## may_dummy      1.189e+01  1.348e+00   8.822 8.56e-16 ***
## jun_dummy      1.226e+01  2.155e+00   5.686 5.02e-08 ***
## jul_dummy      1.557e+01  2.737e+00   5.690 4.93e-08 ***
## aug_dummy      1.529e+01  2.457e+00   6.221 3.25e-09 ***
## sep_dummy      6.801e+00  1.782e+00   3.817 0.000185 ***
## oct_dummy             NA         NA      NA       NA
## yr_2018_dummy -1.341e+00  6.311e-01  -2.125 0.034904 *
## yr_2019_dummy -1.237e+00  6.229e-01  -1.986 0.048510 *
## yr_2020_dummy -1.376e+00  6.152e-01  -2.237 0.026488 *
## yr_2021_dummy  2.780e+00  6.168e-01   4.508 1.16e-05 ***
## yr_2022_dummy         NA         NA      NA       NA
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 6.663637)
##
##     Null deviance: 13413.6  on 203  degrees of freedom
## Residual deviance:  1226.1  on 184  degrees of freedom
## AIC: 986.8
##
## Number of Fisher Scoring iterations: 2
```

Linear Model RMSE = 2.45

Splitting the Data Without Dummy Variables - Testing Linear Models: 75% train/test split

```
sample_size = floor(0.75 * nrow(test))
set.seed(09111997)
split_dat = sample(seq_len(nrow(test)), size = sample_size, replace=FALSE)

ozone_train = test[split_dat, ]
ozone_test = test[-split_dat, ]

lm.final = glm(mda8~., data = ozone_train)
pred.vals = predict(lm.final, new_data=ozone_test)
summary(lm.final)
```

```
##
## Call:
## glm(formula = mda8 ~ ., data = ozone_train)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.2921   -2.6622   -0.0446    3.2560   13.6747
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.342e+02  2.371e+01 -14.097  < 2e-16 ***
## ndvi         7.682e+00  7.243e+00   1.061    0.290
## elev         7.218e-02  6.667e-03  10.827  < 2e-16 ***
## dist2road   -7.165e-04  1.505e-04  -4.762 3.71e-06 ***
## road_length -5.763e-04  3.900e-04  -1.478    0.141
## tmax         8.335e-01  5.860e-02  14.224  < 2e-16 ***
## rhmax        1.107e-01  7.404e-02   1.495    0.136
## pmax         6.934e-01  5.925e-01   1.170    0.243
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 25.25761)
##
##     Null deviance: 13413.6  on 203  degrees of freedom
## Residual deviance:  4950.5  on 196  degrees of freedom
## AIC: 1247.5
##
## Number of Fisher Scoring iterations: 2
```

Linear Model RMSE = 4.93