Project1.R

adnan

2022-11-01

library('tidyverse')

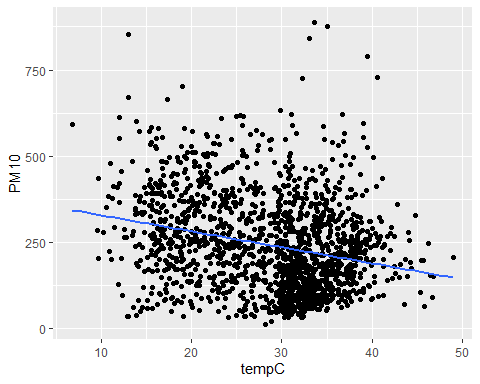
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.3.6 ✔ purrr 0.3.4   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.4.1   
## ✔ readr 2.1.2 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library('ggplot2')  
library('heatmaply')

## Loading required package: plotly  
##   
## Attaching package: 'plotly'  
##   
## The following object is masked from 'package:ggplot2':  
##   
## last\_plot  
##   
## The following object is masked from 'package:stats':  
##   
## filter  
##   
## The following object is masked from 'package:graphics':  
##   
## layout  
##   
## Loading required package: viridis  
## Loading required package: viridisLite  
##   
## ======================  
## Welcome to heatmaply version 1.4.0  
##   
## Type citation('heatmaply') for how to cite the package.  
## Type ?heatmaply for the main documentation.  
##   
## The github page is: https://github.com/talgalili/heatmaply/  
## Please submit your suggestions and bug-reports at: https://github.com/talgalili/heatmaply/issues  
## You may ask questions at stackoverflow, use the r and heatmaply tags:   
## https://stackoverflow.com/questions/tagged/heatmaply  
## ======================

Delhi <- read.csv('D:\\Adnan\\Predictive Analytics Course\\Multivariate Analysis\\Project 1\\Data\\AirQualityIndia\\2015-2020\\delhi.csv\\DelhiAQI.csv')  
Delhi\_24H <- read.csv('D:\\Adnan\\Predictive Analytics Course\\Multivariate Analysis\\Project 1\\Data\\AirQualityIndia\\2015-2020\\delhi.csv\\DelhiAQI\_24H.csv')  
  
Delhi\_24H %>% ggplot(aes(x=tempC, y = PM10)) + geom\_point() + geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'

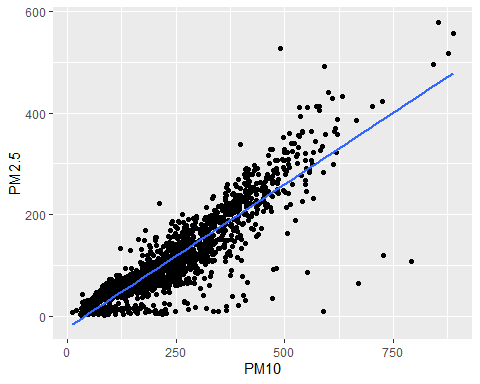


cor(Delhi$PM10, Delhi$tempC)

## [1] -0.216062

#With the above negative trend of -0.216062 shows that, decrease in temperature leads yo increase in particular matter PM10.  
  
Delhi\_24H %>% ggplot(aes(x= PM10, y = PM2.5)) + geom\_point() + geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



cor(Delhi$PM10, Delhi$PM2.5)

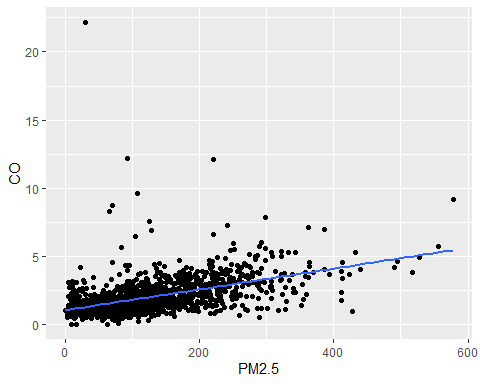
## [1] 0.8428164

Delhi\_24H %>% ggplot(aes(x = PM2.5, y=CO)) + geom\_point() + geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 36 rows containing non-finite values (stat\_smooth).

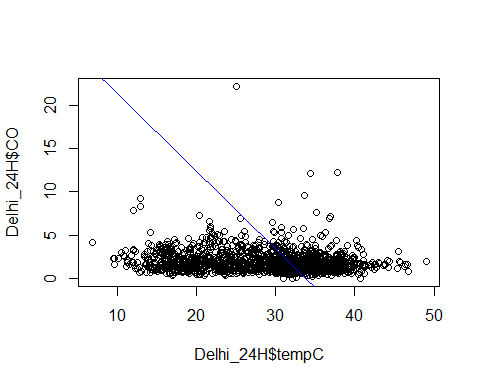
## Warning: Removed 36 rows containing missing values (geom\_point).



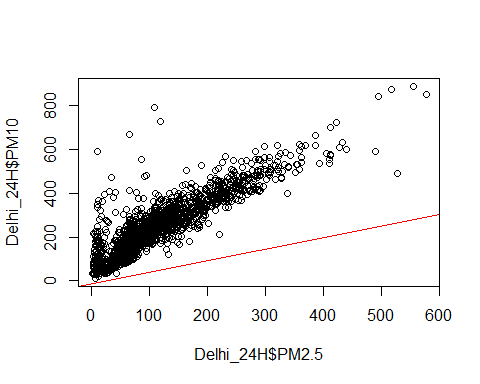
cor(Delhi$PM2.5, Delhi$humidity)

## [1] -0.1384396

plot(x=Delhi\_24H$tempC, y=Delhi\_24H$CO)  
abline(lm(tempC ~ CO, data = Delhi\_24H), col = 'blue')



PM <- lm(PM2.5~PM10, data = Delhi)  
plot(x = Delhi\_24H$PM2.5, y = Delhi\_24H$PM10, abline(PM, col = 'red'))



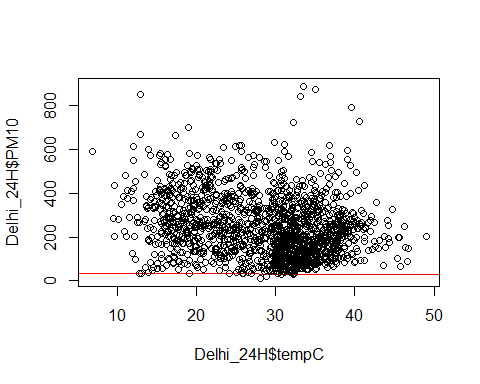
coef(PM)

## (Intercept) PM10   
## -11.9004325 0.5239862

predict(PM, data.frame(PM10 = c(100)))

## 1   
## 40.49819

tempoo <-lm(tempC ~ PM10, data = Delhi\_24H)  
plot(x = Delhi\_24H$tempC, y = Delhi\_24H$PM10, abline(tempoo, col = 'red'))

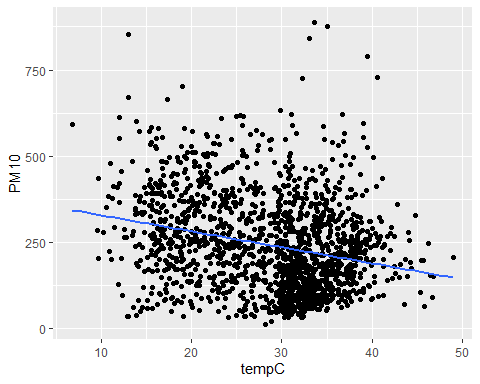


coef(tempoo)

## (Intercept) PM10   
## 32.03996467 -0.01404597

Delhi\_24H %>% ggplot(aes(x=tempC, y = PM10)) + geom\_point() + geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'

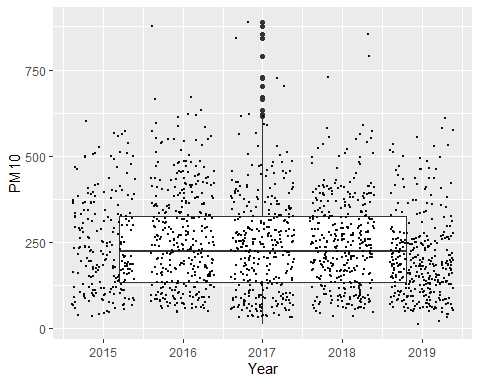


cor(Delhi\_24H$tempC, Delhi\_24H$PM10)

## [1] -0.2554885

Delhi\_24H %>% ggplot(aes(x=Year, y=PM10)) + geom\_boxplot() + geom\_jitter(color = 'black', size = 0.4, alpha=0.9)

## Warning: Continuous x aesthetic -- did you forget aes(group=...)?



Trial1\_Model <- lm(tempC ~ PM10, data = Delhi)  
coef(Trial1\_Model)

## (Intercept) PM10   
## 31.13682665 -0.01057238

Trial2\_Model <- lm(tempC ~ PM10 + PM2.5 + CO, data = Delhi)  
coef(Trial2\_Model)

## (Intercept) PM10 PM2.5 CO   
## 30.920792710 0.006902057 -0.032203228 -0.156366798

summary(Trial1\_Model)

##   
## Call:  
## lm(formula = tempC ~ PM10, data = Delhi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.2593 -5.3739 0.7543 5.4764 22.6109   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.1368267 0.0749681 415.33 <2e-16 \*\*\*  
## PM10 -0.0105724 0.0002577 -41.03 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.887 on 34378 degrees of freedom  
## Multiple R-squared: 0.04668, Adjusted R-squared: 0.04666   
## F-statistic: 1683 on 1 and 34378 DF, p-value: < 2.2e-16

summary(Trial2\_Model)

##   
## Call:  
## lm(formula = tempC ~ PM10 + PM2.5 + CO, data = Delhi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -30.936 -5.088 0.695 5.258 36.743   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 30.9207927 0.0766201 403.560 < 2e-16 \*\*\*  
## PM10 0.0069021 0.0004838 14.266 < 2e-16 \*\*\*  
## PM2.5 -0.0322032 0.0007754 -41.530 < 2e-16 \*\*\*  
## CO -0.1563668 0.0211148 -7.406 1.34e-13 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.632 on 32079 degrees of freedom  
## (2297 observations deleted due to missingness)  
## Multiple R-squared: 0.09865, Adjusted R-squared: 0.09856   
## F-statistic: 1170 on 3 and 32079 DF, p-value: < 2.2e-16

Full\_ModelT <- lm(AQI\_VALUE ~ PM10 + PM2.5 + SO2 + sunHour + mintempC + humidity + windspeedKmph + Hour, data = Delhi)  
coef(Full\_ModelT)

## (Intercept) PM10 PM2.5 SO2 sunHour   
## 239.5286124 0.3274255 0.4151534 0.1879108 -3.7466869   
## mintempC humidity windspeedKmph Hour   
## -1.6321282 -1.0531051 0.3802935 -0.9397992

summary(Full\_ModelT)

##   
## Call:  
## lm(formula = AQI\_VALUE ~ PM10 + PM2.5 + SO2 + sunHour + mintempC +   
## humidity + windspeedKmph + Hour, data = Delhi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -446.85 -50.60 -14.62 46.26 599.93   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 239.528612 3.124466 76.662 < 2e-16 \*\*\*  
## PM10 0.327425 0.005437 60.217 < 2e-16 \*\*\*  
## PM2.5 0.415153 0.008432 49.238 < 2e-16 \*\*\*  
## SO2 0.187911 0.026752 7.024 2.2e-12 \*\*\*  
## sunHour -3.746687 0.286459 -13.079 < 2e-16 \*\*\*  
## mintempC -1.632128 0.084926 -19.218 < 2e-16 \*\*\*  
## humidity -1.053105 0.025620 -41.105 < 2e-16 \*\*\*  
## windspeedKmph 0.380294 0.099591 3.819 0.000134 \*\*\*  
## Hour -0.939799 0.065335 -14.384 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 79.8 on 32905 degrees of freedom  
## (1466 observations deleted due to missingness)  
## Multiple R-squared: 0.6372, Adjusted R-squared: 0.6371   
## F-statistic: 7223 on 8 and 32905 DF, p-value: < 2.2e-16

FullModelT2 <- lm(tempC ~ AQI\_VALUE + Hour + sunHour + humidity + uvIndex + Main\_Pollutant + windspeedKmph + mintempC, data = Delhi)  
coef(FullModelT2)

## (Intercept) AQI\_VALUE Hour sunHour   
## 6.522022239 -0.002191353 0.090440997 0.320634671   
## humidity uvIndex Main\_PollutantNO2 Main\_PollutantOzone   
## -0.105786316 -0.133477458 0.682008260 0.717048992   
## Main\_PollutantPM10 Main\_PollutantPM2.5 windspeedKmph mintempC   
## 0.431246316 0.571476573 0.110187178 0.884862502

summary(FullModelT2)

##   
## Call:  
## lm(formula = tempC ~ AQI\_VALUE + Hour + sunHour + humidity +   
## uvIndex + Main\_Pollutant + windspeedKmph + mintempC, data = Delhi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -38.514 -1.982 -0.068 1.887 10.154   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 6.5220222 0.1948702 33.469 < 2e-16 \*\*\*  
## AQI\_VALUE -0.0021914 0.0001297 -16.895 < 2e-16 \*\*\*  
## Hour 0.0904410 0.0022514 40.171 < 2e-16 \*\*\*  
## sunHour 0.3206347 0.0102348 31.328 < 2e-16 \*\*\*  
## humidity -0.1057863 0.0010341 -102.300 < 2e-16 \*\*\*  
## uvIndex -0.1334775 0.0330336 -4.041 5.34e-05 \*\*\*  
## Main\_PollutantNO2 0.6820083 0.2382247 2.863 0.00420 \*\*   
## Main\_PollutantOzone 0.7170490 0.1811393 3.959 7.56e-05 \*\*\*  
## Main\_PollutantPM10 0.4312463 0.1311265 3.289 0.00101 \*\*   
## Main\_PollutantPM2.5 0.5714766 0.1324046 4.316 1.59e-05 \*\*\*  
## windspeedKmph 0.1101872 0.0034738 31.719 < 2e-16 \*\*\*  
## mintempC 0.8848625 0.0063397 139.575 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.842 on 34348 degrees of freedom  
## (20 observations deleted due to missingness)  
## Multiple R-squared: 0.8762, Adjusted R-squared: 0.8762   
## F-statistic: 2.21e+04 on 11 and 34348 DF, p-value: < 2.2e-16

FullModelT3 <- lm(AQI\_VALUE ~ mintempC + maxtempC + PM10 + PM2.5 + sunHour + humidity + uvIndex + Month, data = Delhi)  
coef(FullModelT3)

## (Intercept) mintempC maxtempC PM10 PM2.5   
## 234.989527300 -1.347309494 0.232556859 0.331214780 0.440920161   
## sunHour humidity uvIndex Month   
## -3.916175964 -1.036725477 -2.038028716 -0.001325588

summary(FullModelT3)

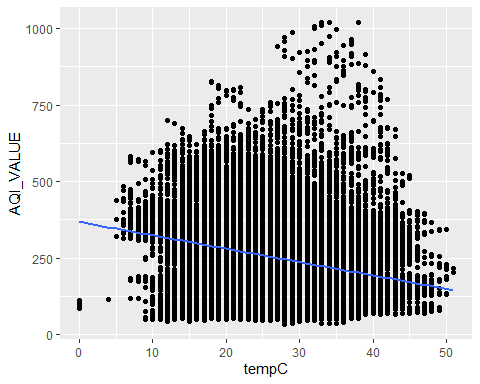
##   
## Call:  
## lm(formula = AQI\_VALUE ~ mintempC + maxtempC + PM10 + PM2.5 +   
## sunHour + humidity + uvIndex + Month, data = Delhi)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -469.65 -51.82 -14.46 47.85 610.88   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 234.989527 5.017211 46.837 < 2e-16 \*\*\*  
## mintempC -1.347309 0.220935 -6.098 1.08e-09 \*\*\*  
## maxtempC 0.232557 0.250727 0.928 0.354   
## PM10 0.331215 0.005202 63.668 < 2e-16 \*\*\*  
## PM2.5 0.440920 0.008240 53.512 < 2e-16 \*\*\*  
## sunHour -3.916176 0.352784 -11.101 < 2e-16 \*\*\*  
## humidity -1.036725 0.031903 -32.496 < 2e-16 \*\*\*  
## uvIndex -2.038029 1.066366 -1.911 0.056 .   
## Month -0.001326 0.167885 -0.008 0.994   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 82.09 on 34351 degrees of freedom  
## (20 observations deleted due to missingness)  
## Multiple R-squared: 0.6348, Adjusted R-squared: 0.6347   
## F-statistic: 7463 on 8 and 34351 DF, p-value: < 2.2e-16

Delhi %>% ggplot(aes(x=tempC, y = AQI\_VALUE)) + geom\_point() + geom\_smooth(method = 'lm', se=FALSE)

## `geom\_smooth()` using formula 'y ~ x'

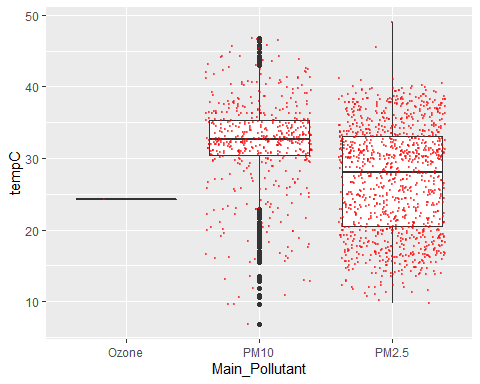
## Warning: Removed 20 rows containing non-finite values (stat\_smooth).

## Warning: Removed 20 rows containing missing values (geom\_point).



Delhi\_24H %>% ggplot(aes(x=Main\_Pollutant, y = tempC)) + geom\_boxplot() + geom\_jitter(color = 'red', size = 0.2, alpha = 0.5) + geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'



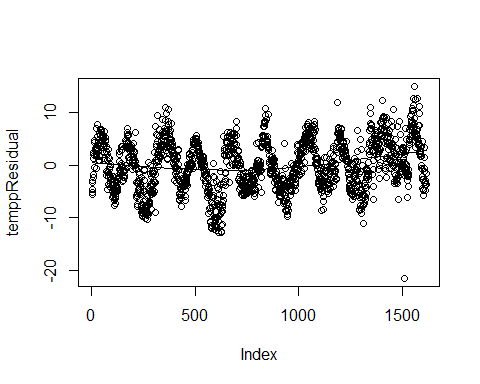
predict(FullModelT2, data.frame(AQI\_VALUE = c(400, 100),  
 Hour = c(20, 13),  
 sunHour = c(10,14),  
 humidity = c(70,21),  
 uvIndex = c(5,8),  
 Main\_Pollutant = c('PM10', 'PM10'),  
 windspeedKmph = c(17,7),  
 mintempC = c(22,33)))

## 1 2   
## 24.35962 39.08119

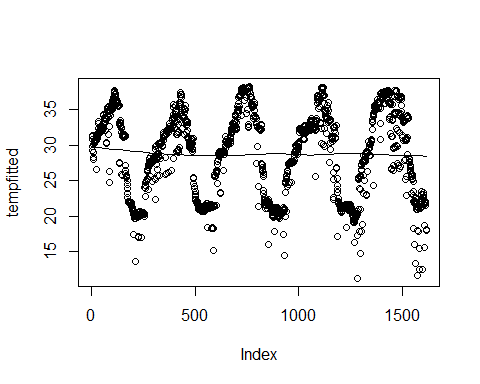
tempp <- lm(tempC ~ Main\_Pollutant + AQI\_VALUE + sunHour, data = Delhi\_24H)  
coef(tempp)

## (Intercept) Main\_PollutantPM2.5 AQI\_VALUE sunHour   
## 14.73081091 -0.26026171 -0.03098497 1.95218705

temppResidual <- tempp$residuals  
tempfitted <- tempp$fitted.values  
scatter.smooth(temppResidual)



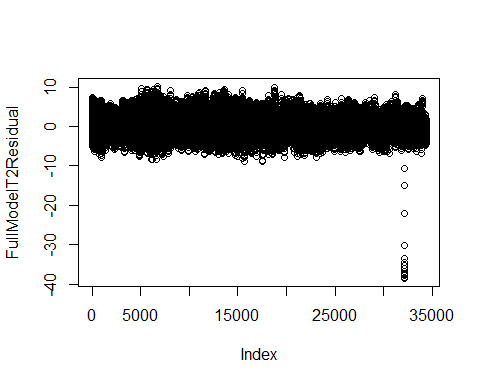
scatter.smooth(tempfitted)



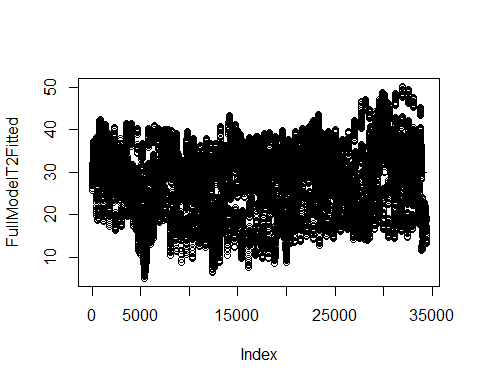
summary(tempp)

##   
## Call:  
## lm(formula = tempC ~ Main\_Pollutant + AQI\_VALUE + sunHour, data = Delhi\_24H)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -21.5101 -3.3917 0.0201 3.6094 15.0563   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.730811 0.931055 15.822 <2e-16 \*\*\*  
## Main\_PollutantPM2.5 -0.260262 0.294693 -0.883 0.377   
## AQI\_VALUE -0.030985 0.001831 -16.922 <2e-16 \*\*\*  
## sunHour 1.952187 0.059690 32.706 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.701 on 1611 degrees of freedom  
## (1 observation deleted due to missingness)  
## Multiple R-squared: 0.6132, Adjusted R-squared: 0.6125   
## F-statistic: 851.4 on 3 and 1611 DF, p-value: < 2.2e-16

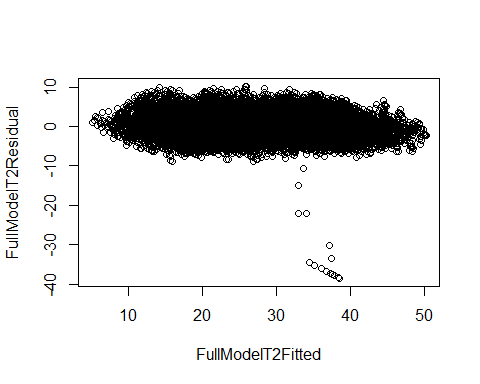
FullModelT2Residual <- FullModelT2$residuals  
FullModelT2Fitted <- FullModelT2$fitted.values  
scatter.smooth(FullModelT2Residual)



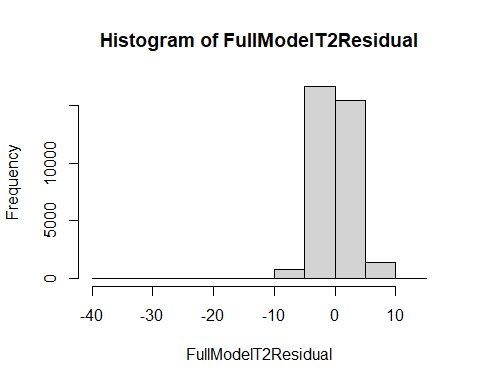
scatter.smooth(FullModelT2Fitted)



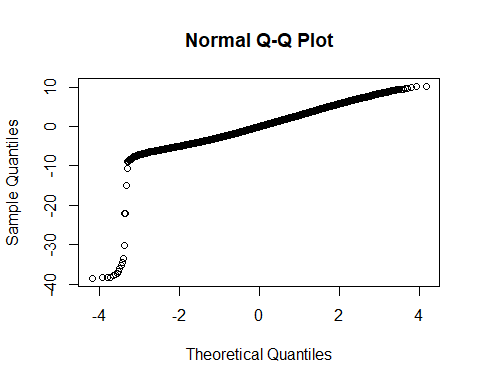
plot(FullModelT2Fitted, FullModelT2Residual)



hist(FullModelT2Residual)



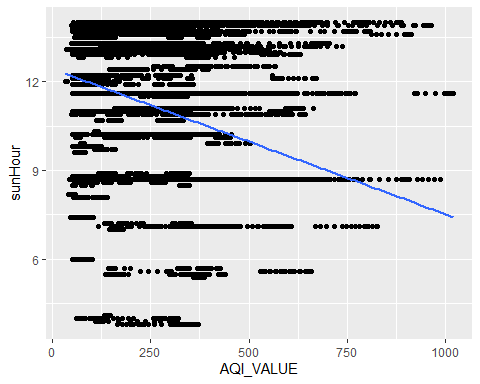
qqnorm(FullModelT2Residual)



Delhi %>% ggplot(aes(x = AQI\_VALUE, y = sunHour)) + geom\_point() + geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 20 rows containing non-finite values (stat\_smooth).  
## Removed 20 rows containing missing values (geom\_point).



Delhi\_24H %>% ggplot(aes(x = sunHour, y = AQI\_VALUE)) + geom\_point() + geom\_smooth(method = 'lm', se = FALSE)

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 1 rows containing non-finite values (stat\_smooth).

## Warning: Removed 1 rows containing missing values (geom\_point).

