Tochi Okorie Dissertation

```r
chooseCRANmirror(graphics=FALSE, ind=1)
knitr::opts_chunk$set(echo = TRUE)
```

Dissertation

Tochi Okorie Carbon footprints of digital systems

```r
pkgs <- c("moments", "ggplot2", "dplyr", "tidyr", "tidyverse")
install.packages(pkgs, repos = "http://cran.us.r-project.org")
```

```
## Installing packages into 'C:/Users/tochi/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'moments' successfully unpacked and MD5 sums checked
## package 'ggplot2' successfully unpacked and MD5 sums checked
## package 'dplyr' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'dplyr'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:
## \Users\tochi\Documents\R\win-library\4.0\00LOCK\dplyr\libs\x64\dplyr.dll to C:
## \Users\tochi\Documents\R\win-library\4.0\dplyr\libs\x64\dplyr.dll: Permission
## denied
```

```
## Warning: restored 'dplyr'
```

```
## package 'tidyr' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'tidyr'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:
## \Users\tochi\Documents\R\win-library\4.0\00LOCK\tidyr\libs\x64\tidyr.dll to C:
## \Users\tochi\Documents\R\win-library\4.0\tidyr\libs\x64\tidyr.dll: Permission
## denied
```

```
## Warning: restored 'tidyr'
```

```
## package 'tidyverse' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\tochi\AppData\Local\Temp\RtmpQRXTPB\downloaded_packages
```

```r
tinytex::install_tinytex()
```

```
## tlmgr conf auxtrees add "C:/PROGRA~1/R/R-40~1.2/share/texmf"
```

```r
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.0.5
```

```r
install.packages("rlang")
```

```
## Installing package into 'C:/Users/tochi/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)
```

```
## package 'rlang' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'rlang'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying C:
## \Users\tochi\Documents\R\win-library\4.0\00LOCK\rlang\libs\x64\rlang.dll to C:
## \Users\tochi\Documents\R\win-library\4.0\rlang\libs\x64\rlang.dll: Permission
## denied
```

```
## Warning: restored 'rlang'
```

```
##
## The downloaded binary packages are in
##  C:\Users\tochi\AppData\Local\Temp\RtmpQRXTPB\downloaded_packages
```

```r
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.5
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## Warning: package 'tibble' was built under R version 4.0.5
```

```
## Warning: package 'tidyr' was built under R version 4.0.5
```

```
## Warning: package 'readr' was built under R version 4.0.3
```

```
## Warning: package 'purrr' was built under R version 4.0.3
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
## Warning: package 'forcats' was built under R version 4.0.5

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
install.packages("lmtest", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/tochi/Documents/R/win-library/4.0'
## (as 'lib' is unspecified)

## package 'lmtest' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\tochi\AppData\Local\Temp\RtmpQRXTPB\downloaded_packages
```

Load the merged carbon footprint data into my notebook.

```r
cf_data <-read.csv("C:/Users/tochi/Desktop/carbonfootprint_data.csv")
```

EXPLORATORY DATA ANALYSIS

Explore the data numerically and graphically. Confirm the variables that are categorical and numerical/continuous and that R has read them in #appropriately

```r
# inspect the dataset
str(cf_data)
```

```
## 'data.frame':    99 obs. of  8 variables:
##  $ WEIGHT_OF_CO2_per_time_visited.grams.: num  1.69 1.48 0.68 1.3 11.87 ...
##  $ GREEN_HOSTING                        : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ WEIGHT_OF_CARBON.In_grams_yearly.    : int  203340 177620 82180 155760 237750 135900 94990 287140
##  $ Energy.Kwh.                          : int  428 374 173 328 2999 316 200 667 1677 444 ...
##  $ Score.percentage.                    : num  0.3 0.35 0.39 0.32 0.39 0.38 0.46 0.32 0.43 0.21 ...
##  $ Google_page_insights                 : int  84 73 79 67 31 88 84 45 53 37 ...
##  $ HTTP_REQUEST                         : int  369 242 105 115 224 102 85 300 130 121 ...
##  $ FINDABILITY.Mozrank.                 : num  9.2 8.4 6.5 5.3 5.5 5.7 5 7.4 5.5 7.7 ...
```

```r
# get a summary report
summary(cf_data)
```

```
##  WEIGHT_OF_CO2_per_time_visited.grams. GREEN_HOSTING
##  Min.   : 0.170                        Min.   :0.0000
##  1st Qu.: 1.330                        1st Qu.:0.0000
##  Median : 2.040                        Median :0.0000
##  Mean   : 2.584                        Mean   :0.1414
##  3rd Qu.: 3.070                        3rd Qu.:0.0000
##  Max.   :12.810                        Max.   :1.0000
##  WEIGHT_OF_CARBON.In_grams_yearly.  Energy.Kwh.     Score.percentage.
##  Min.   : 20560                   Min.   :  43.0   Min.   :0.2000
##  1st Qu.:159265                   1st Qu.: 344.5   1st Qu.:0.3550
```

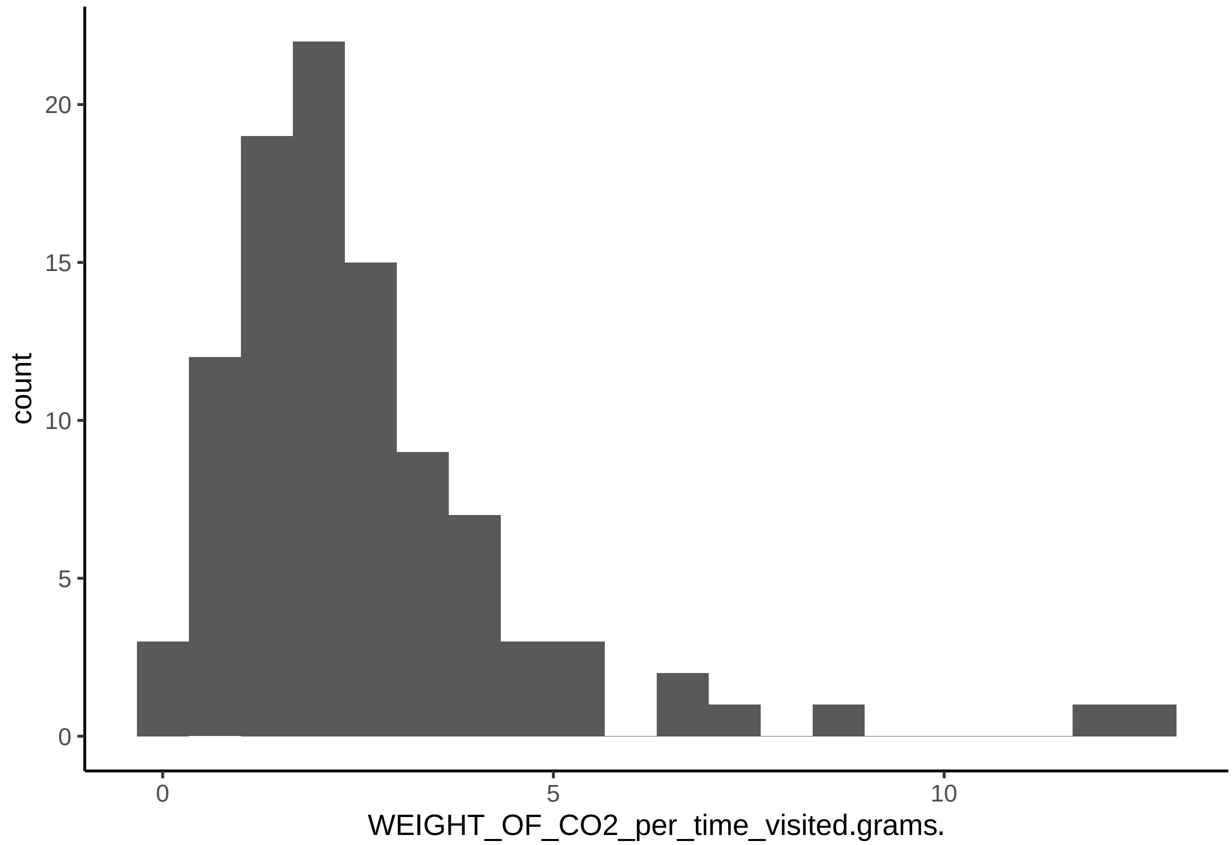
```
##  Median :237750                    Median : 518.0   Median :0.4100
##  Mean   :259377                    Mean   : 661.4   Mean   :0.4058
##  3rd Qu.:329275                    3rd Qu.: 776.0   3rd Qu.:0.4550
##  Max.   :635360                    Max.   :3236.0   Max.   :0.7000
##  Google_page_insights  HTTP_REQUEST   FINDABILITY.Mozrank.
##  Min.   : 14.00       Min.   : 20.0   Min.   :1.300
##  1st Qu.: 52.50       1st Qu.: 85.0   1st Qu.:5.200
##  Median : 71.00       Median :128.0   Median :6.000
##  Mean   : 66.18       Mean   :153.8   Mean   :5.944
##  3rd Qu.: 84.00       3rd Qu.:200.0   3rd Qu.:6.800
##  Max.   :132.00       Max.   :535.0   Max.   :9.200
```

The variable "Greenhosting" should be a categorical variable (actually binary as it only has two levels). R has read it in as numerical so this can be fixed by making it into a Factor.

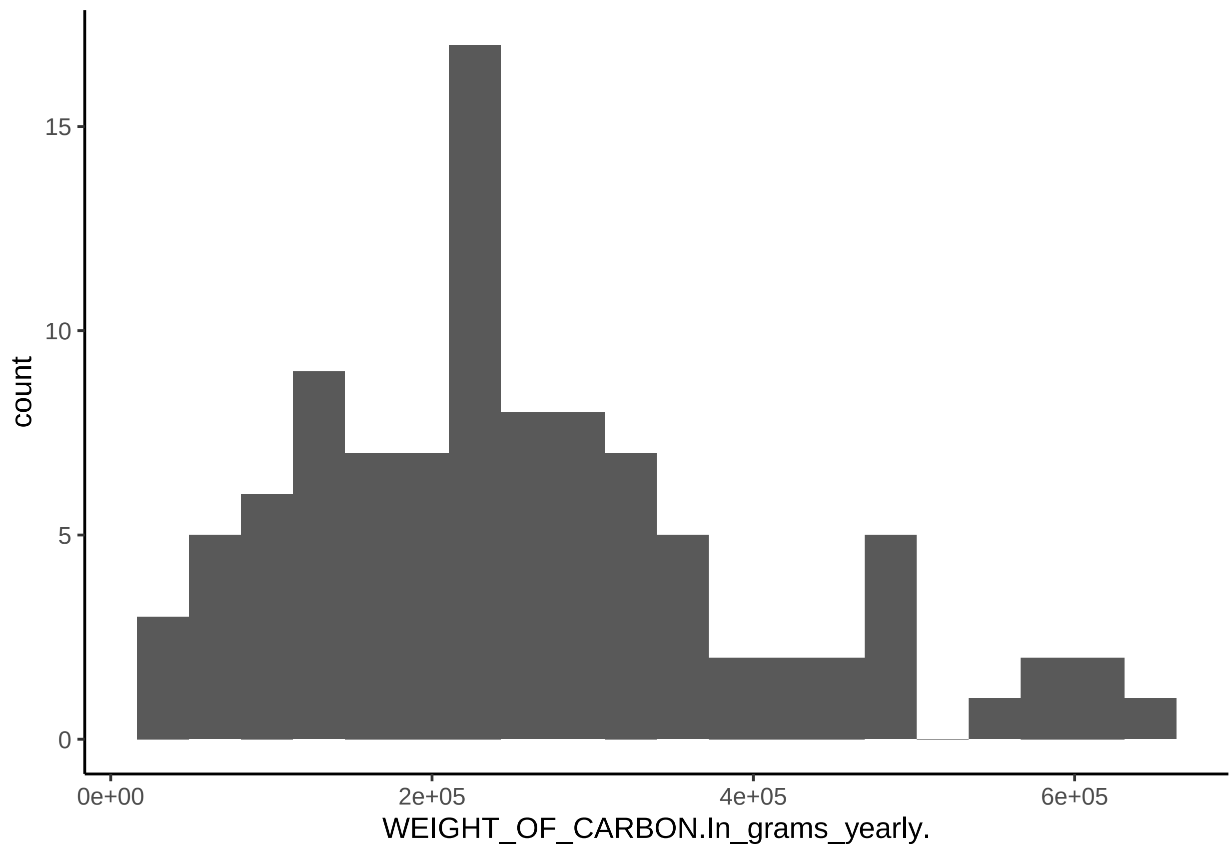
```
#cf_data$GREEN_HOSTING<-as.factor(cf_data$GREEN_HOSTING)
```

Then i look at the distribution of the variables:

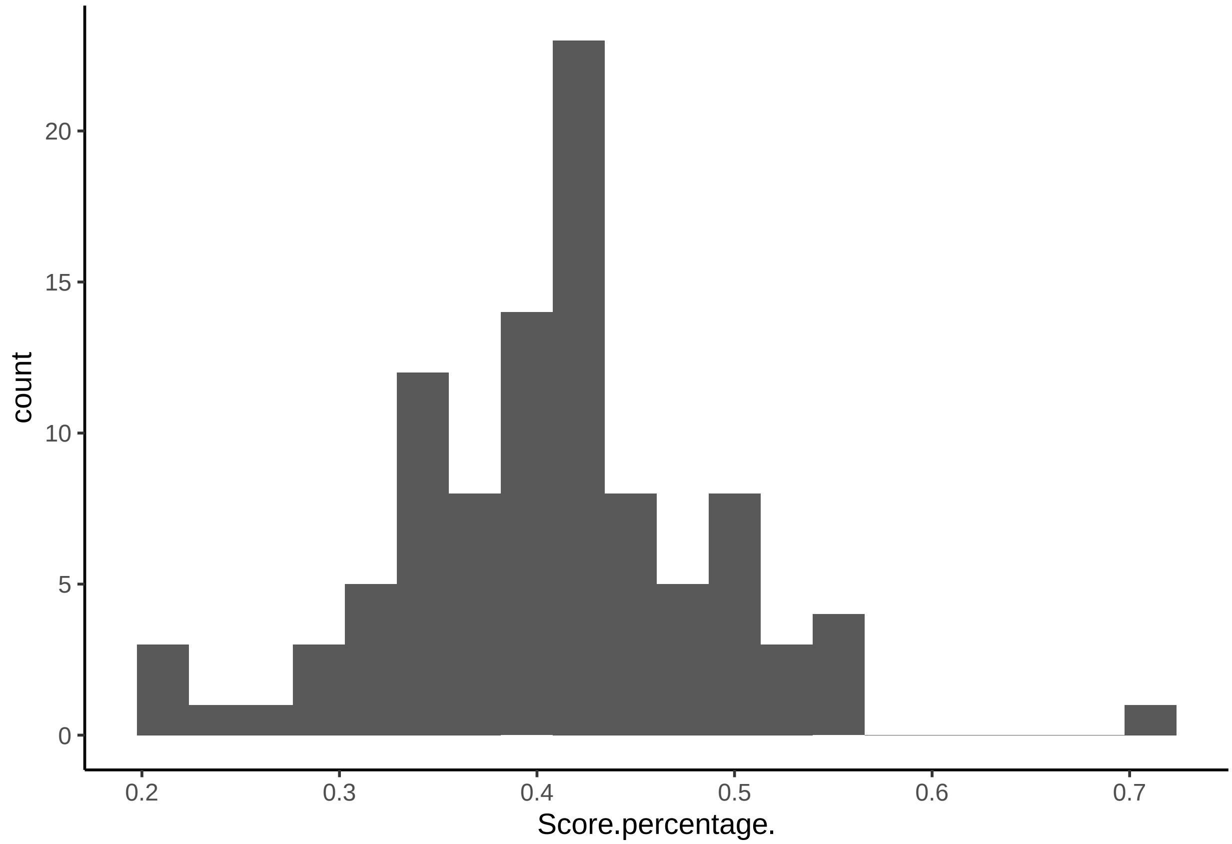
```
ggplot(data = cf_data, aes(x=WEIGHT_OF_CO2_per_time_visited.grams.)) + geom_histogram(bins = 20) + theme
```



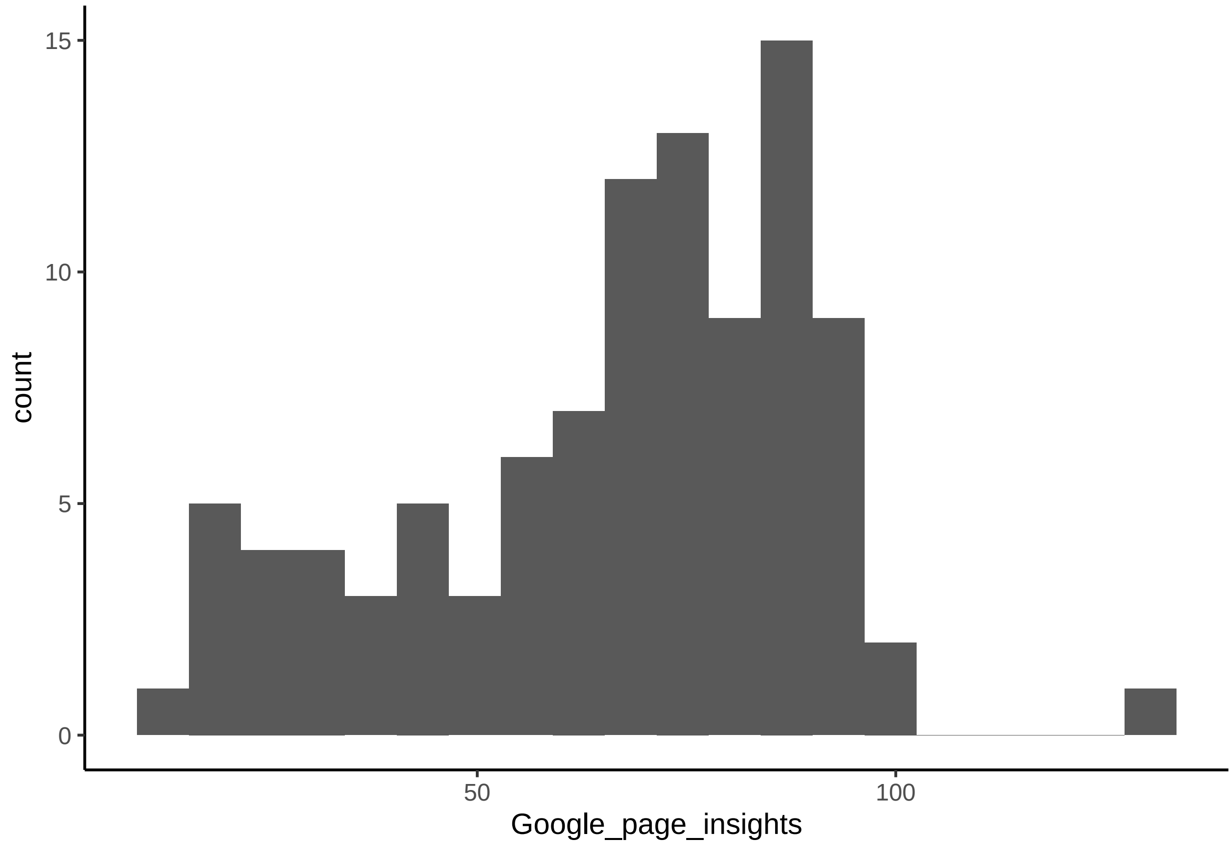

```
ggplot(data = cf_data, aes(x=WEIGHT_OF_CARBON.In_grams_yearly.
)) + geom_histogram(bins = 20) + theme_classic()
```



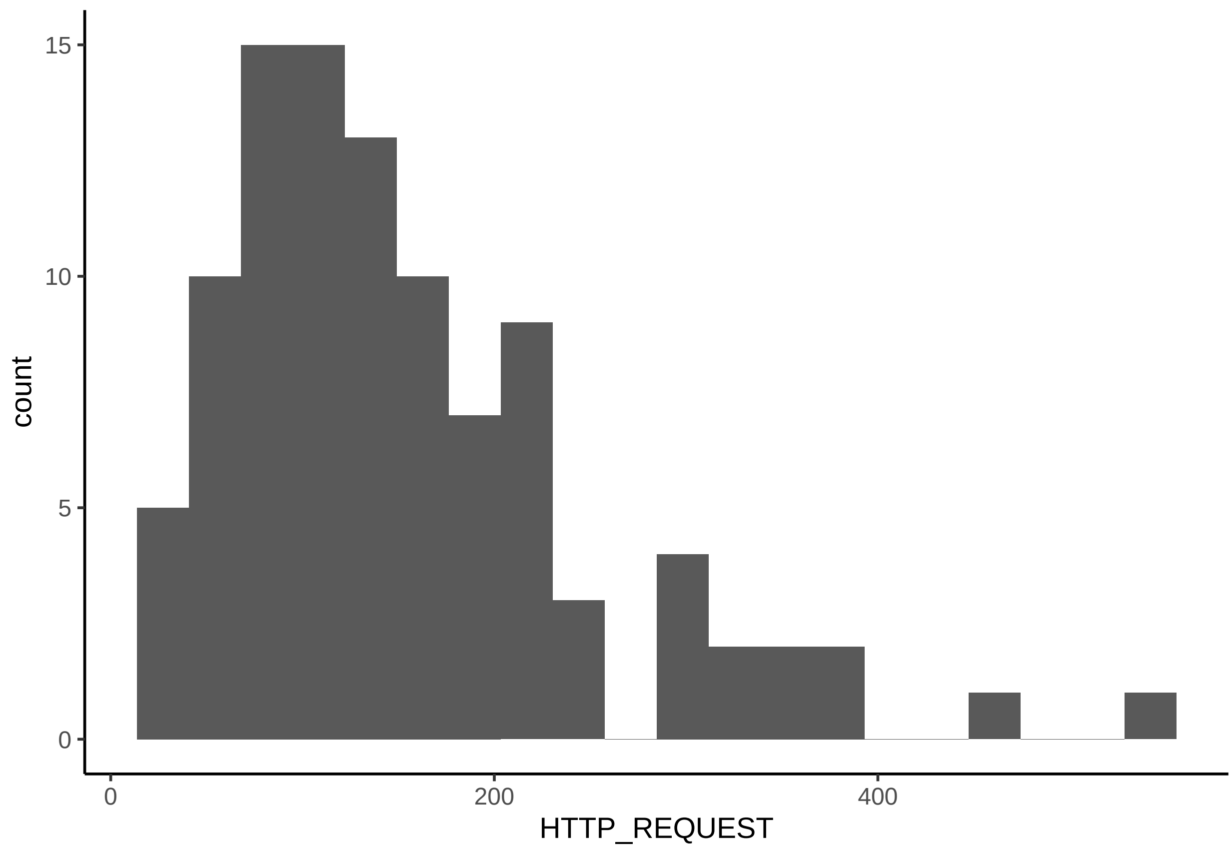

```r
ggplot(data = cf_data, aes(x=Score.percentage.)) + geom_histogram(bins = 20) + theme_classic()
```



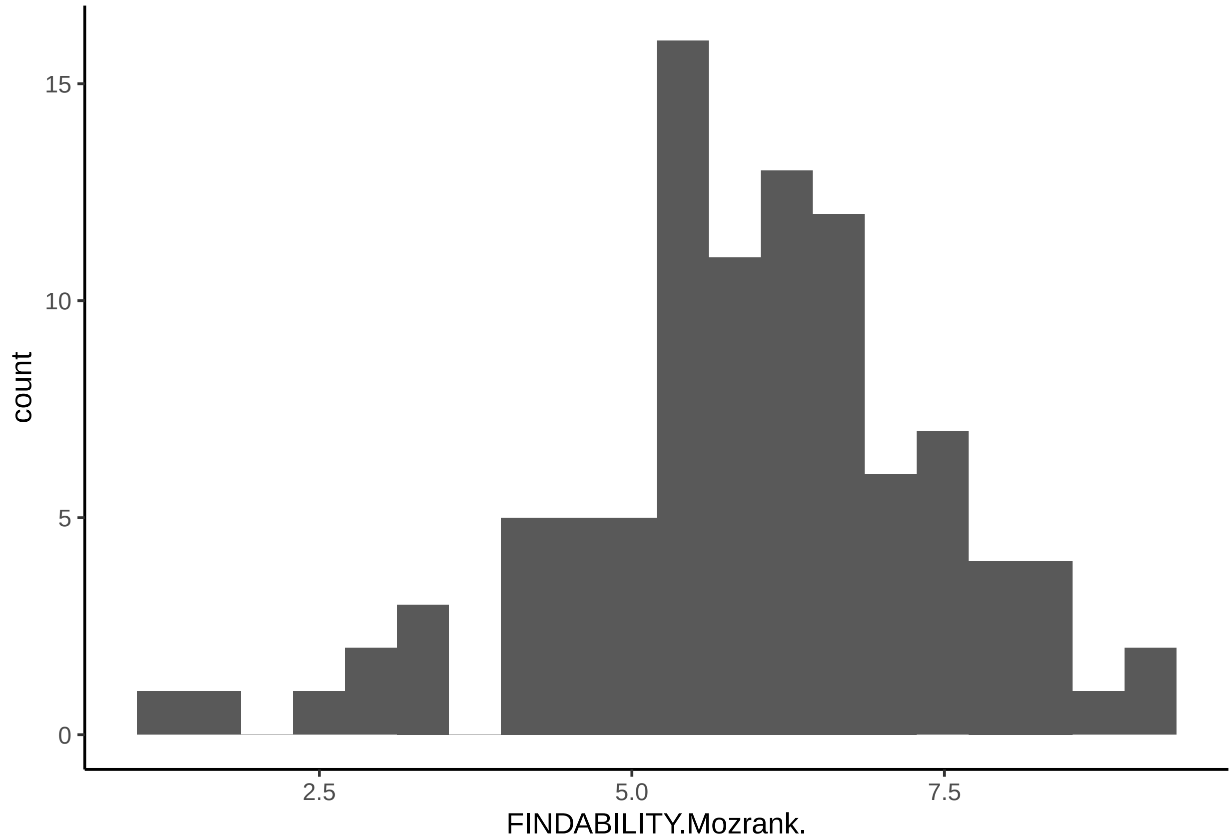

```
ggplot(data = cf_data, aes(x=Google_page_insights)) + geom_histogram(bins = 20) + theme_classic()
```




```r
ggplot(data = cf_data, aes(x=HTTP_REQUEST)) + geom_histogram(bins = 20) + theme_classic()
```

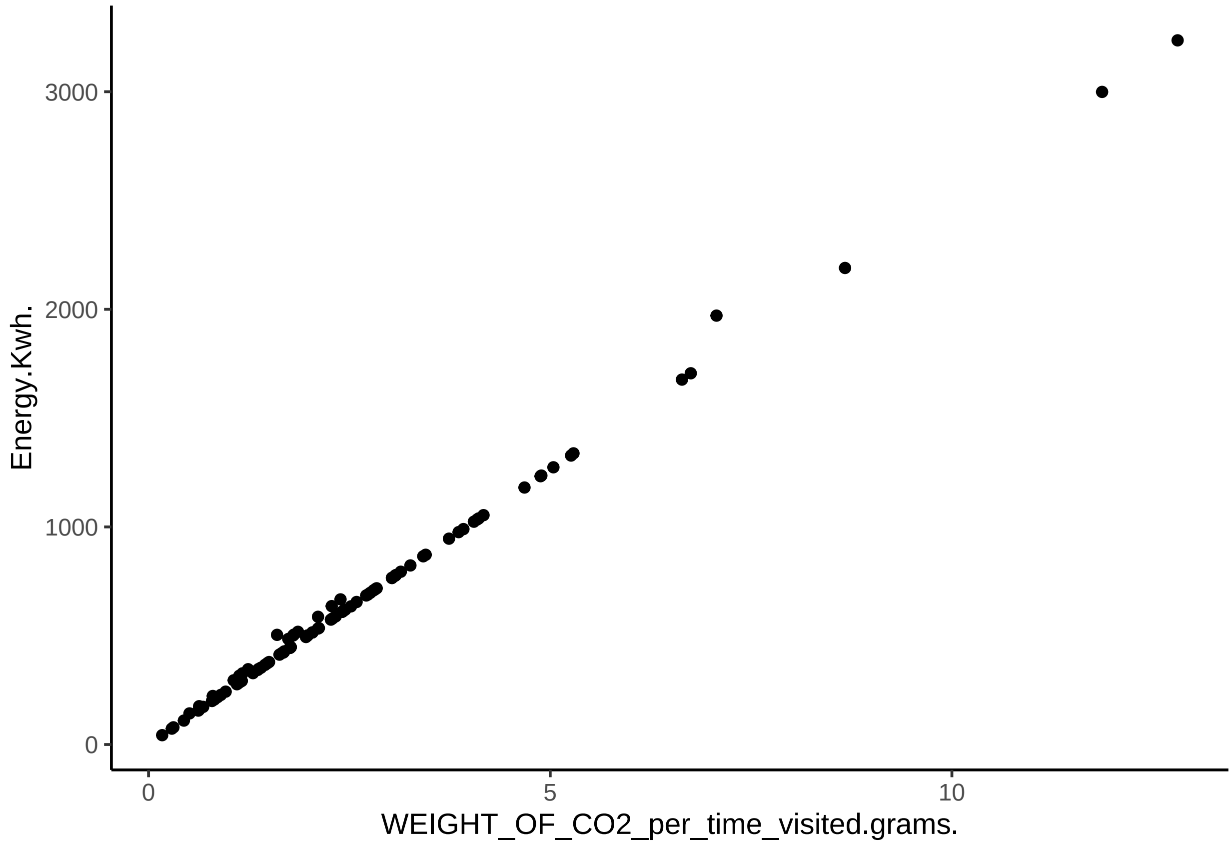



```
ggplot(data = cf_data, aes(x=FINDABILITY.Mozrank.)) + geom_histogram(bins = 20) + theme_classic()
```

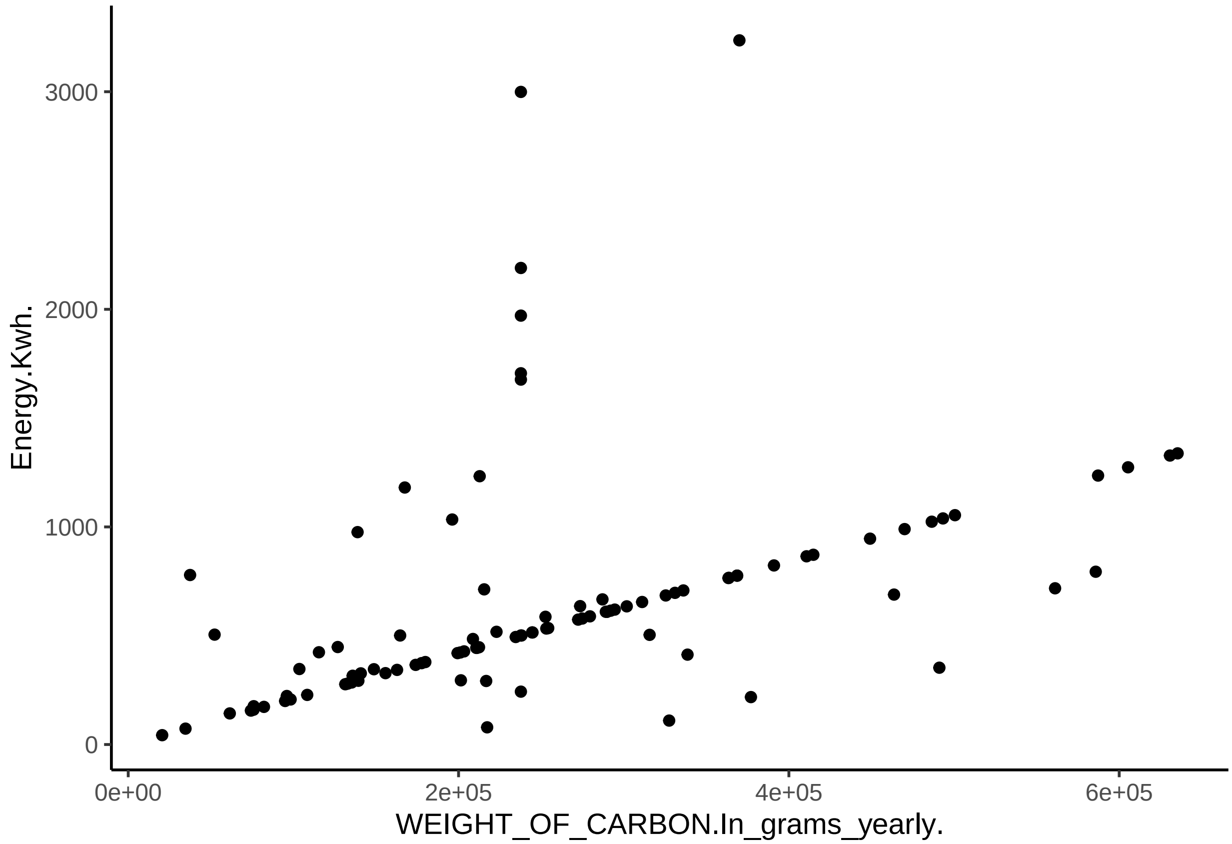


The distribution for Weight per time visited and weight of co2 yearly seems skewed to the left, other variables look generally symmetric. This does not warrant any transformations at this stage.

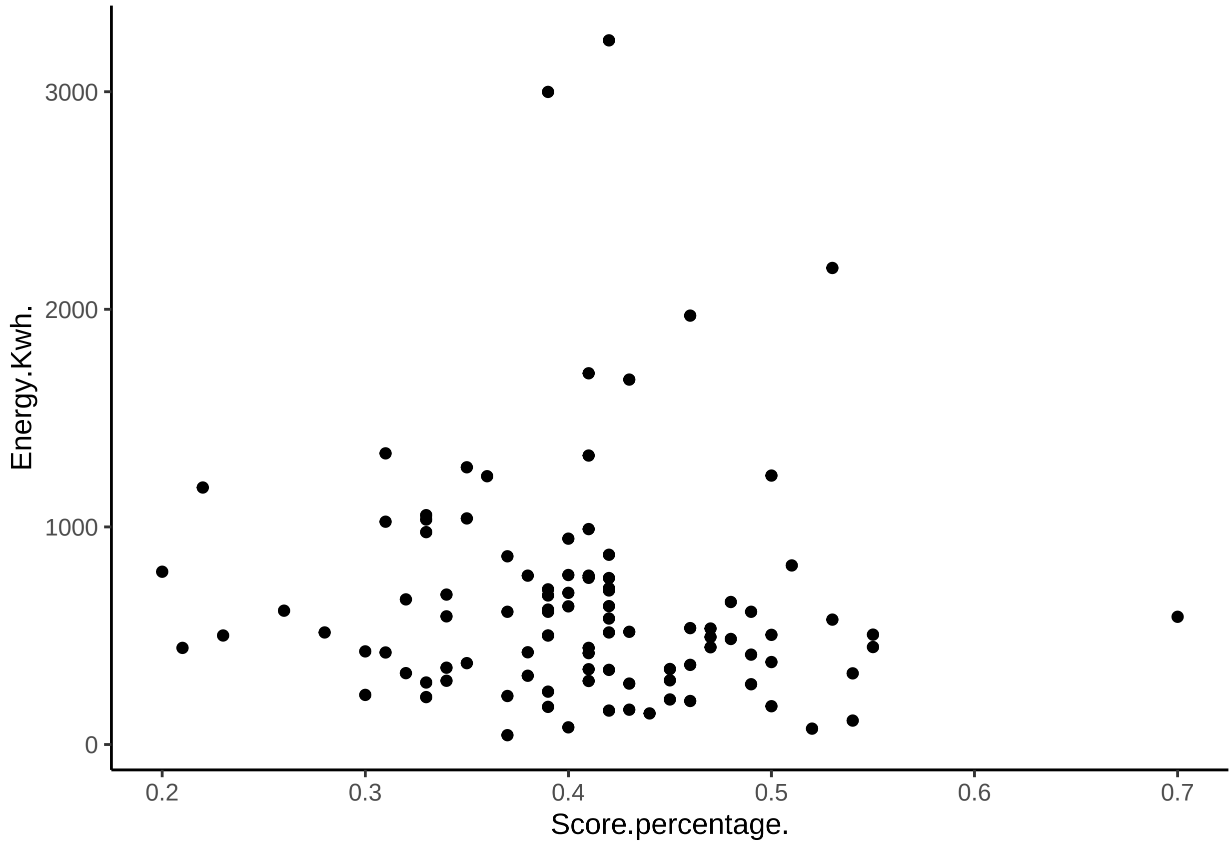
```
ggplot(data = cf_data, aes(x=WEIGHT_OF_CO2_per_time_visited.grams., y=Energy.Kwh.)) + geom_point() + the
```



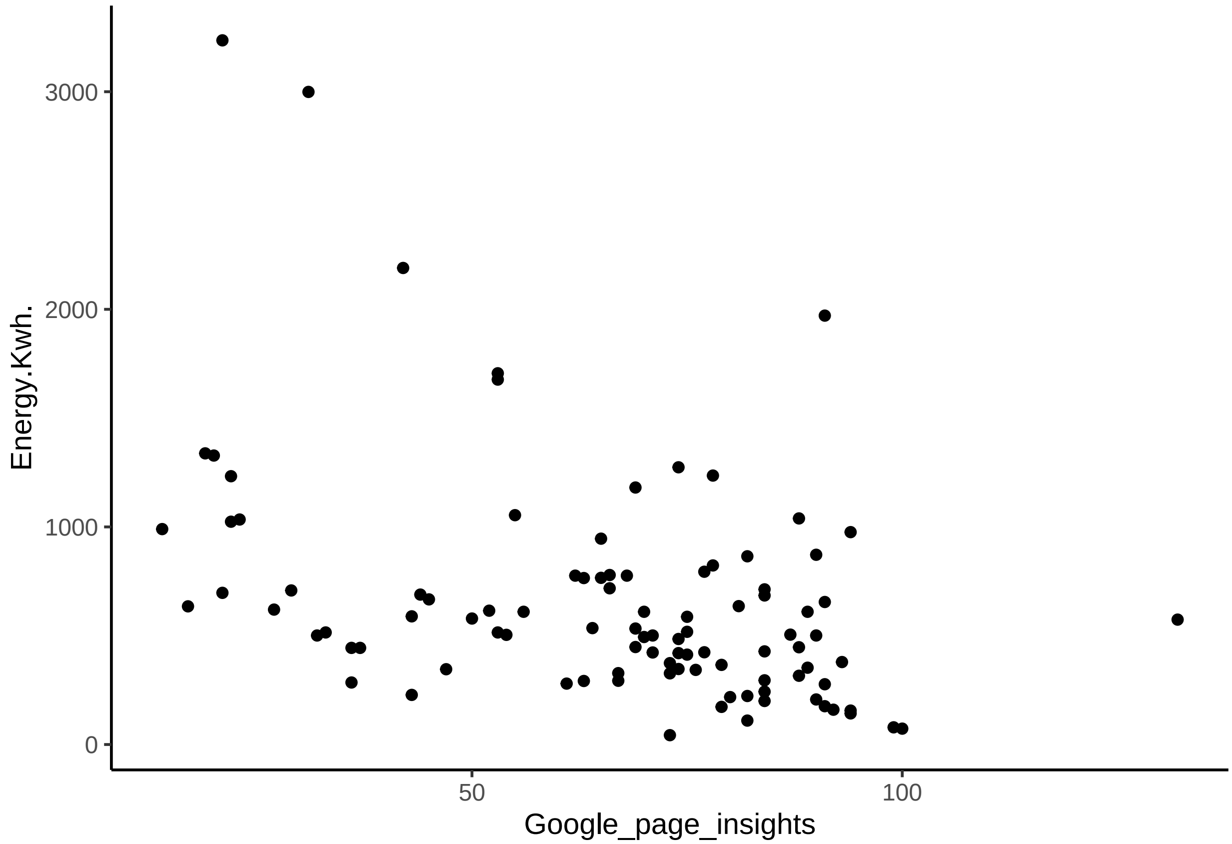

```
ggplot(data = cf_data, aes(x=WEIGHT_OF_CARBON.In_grams_yearly., y=Energy.Kwh.)) + geom_point() + theme_
```



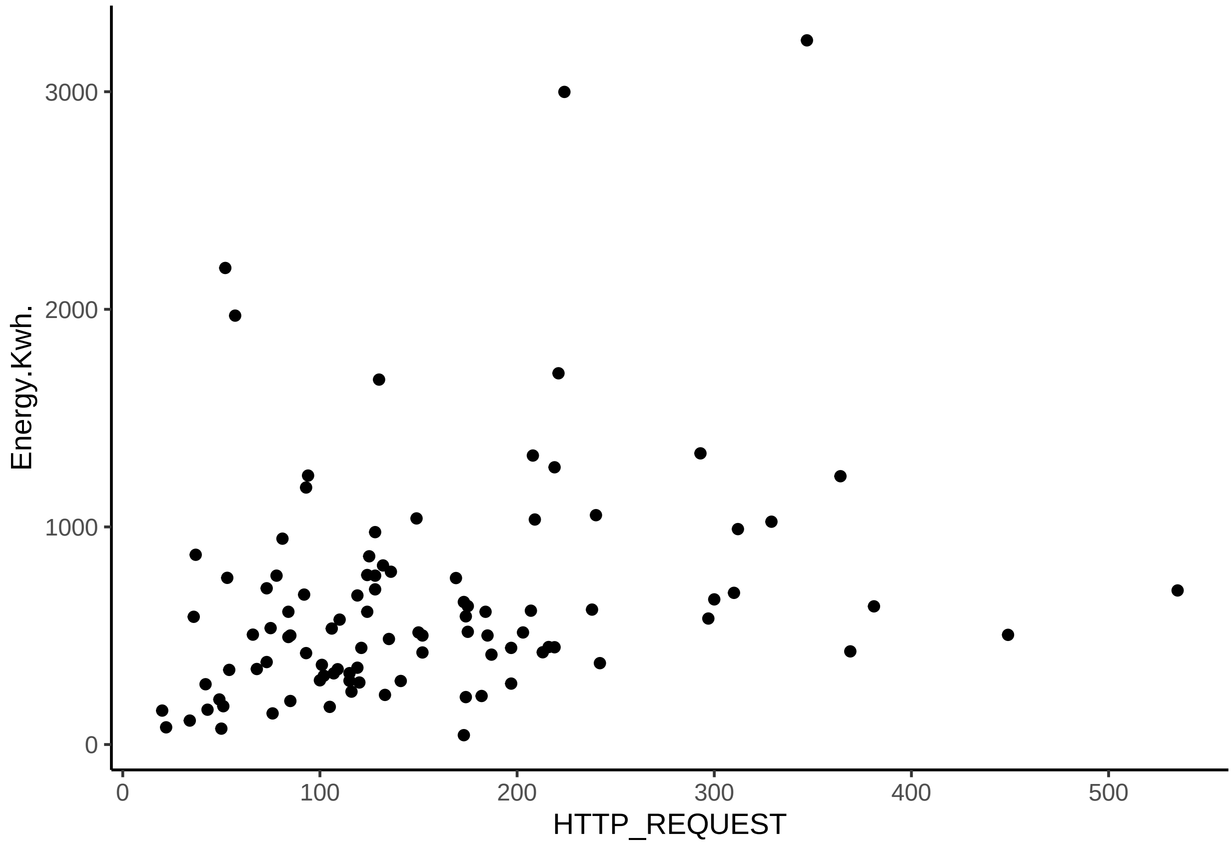

```r
ggplot(data = cf_data, aes(x=Score.percentage., y=Energy.Kwh.)) + geom_point()  + theme_classic()
```



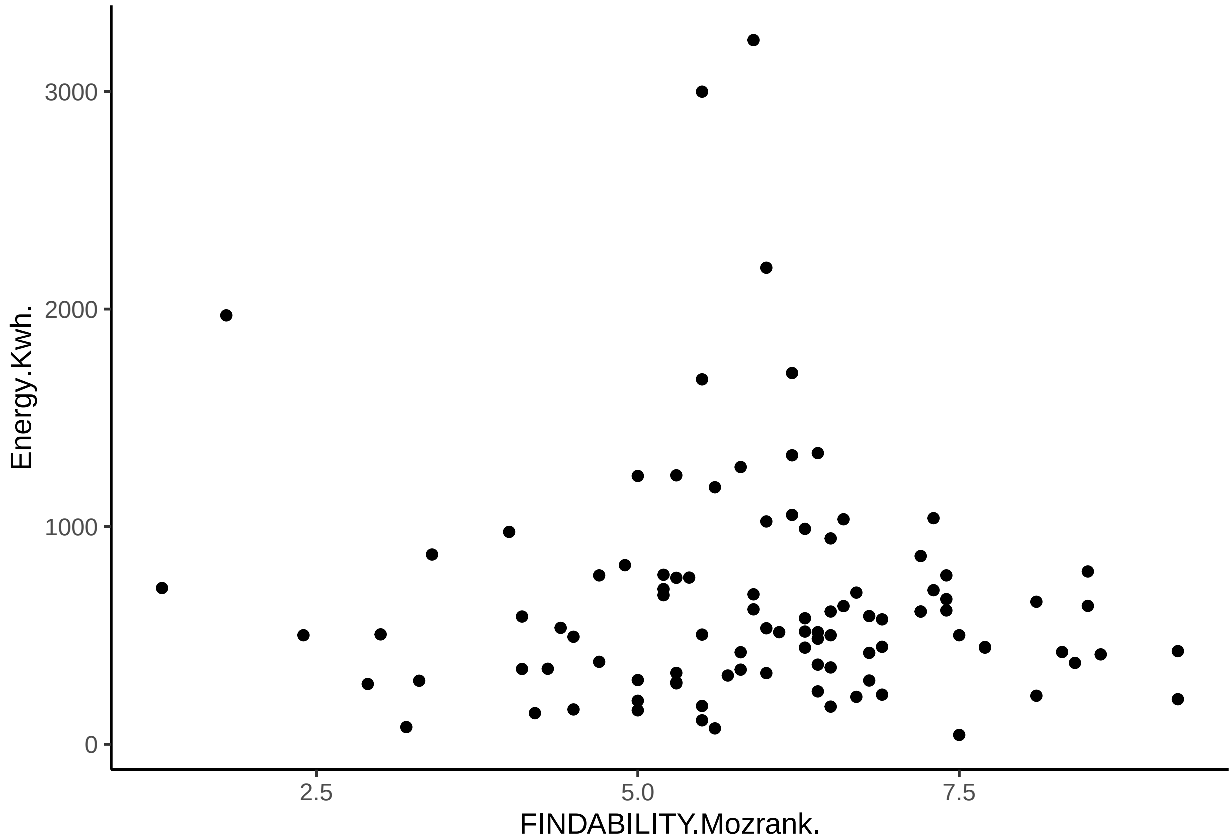

```
ggplot(data = cf_data, aes(x=Google_page_insights, y=Energy.Kwh.)) + geom_point()  + theme_classic()
```




```
ggplot(data = cf_data, aes(x=HTTP_REQUEST, y=Energy.Kwh.)) + geom_point()  + theme_classic()
```




```
ggplot(data = cf_data, aes(x=FINDABILITY.Mozrank., y=Energy.Kwh.)) + geom_point()  + theme_classic()
```




```
#ggplot(cf_data, aes(x=GREEN_HOSTING, y=Energy.Kwh.)) + geom_boxplot()
```

The first two graphs appear to have a linear relationship while the others have no specific pattern just clusters at different regions of the graphs.The collection of scatter plots do not show that most of the variables is clearly linear, but some show a linear trend.

UNSUPERVISED LEARNING

Using unsupervised learning method Principal component analysis:

```
# perform PCA on the cf_data dataset
#   note: variables are centered and scaled before analysis
pc_cf_data <- prcomp(cf_data, center = T, scale. = T)

# inspect the attributes of the PCA object returned by prcomp
attributes(pc_cf_data)
```

```
## $names
## [1] "sdev"     "rotation" "center"   "scale"    "x"
##
## $class
## [1] "prcomp"
```

Visual analysis of PCA results{#Visual_analysis_PCA}

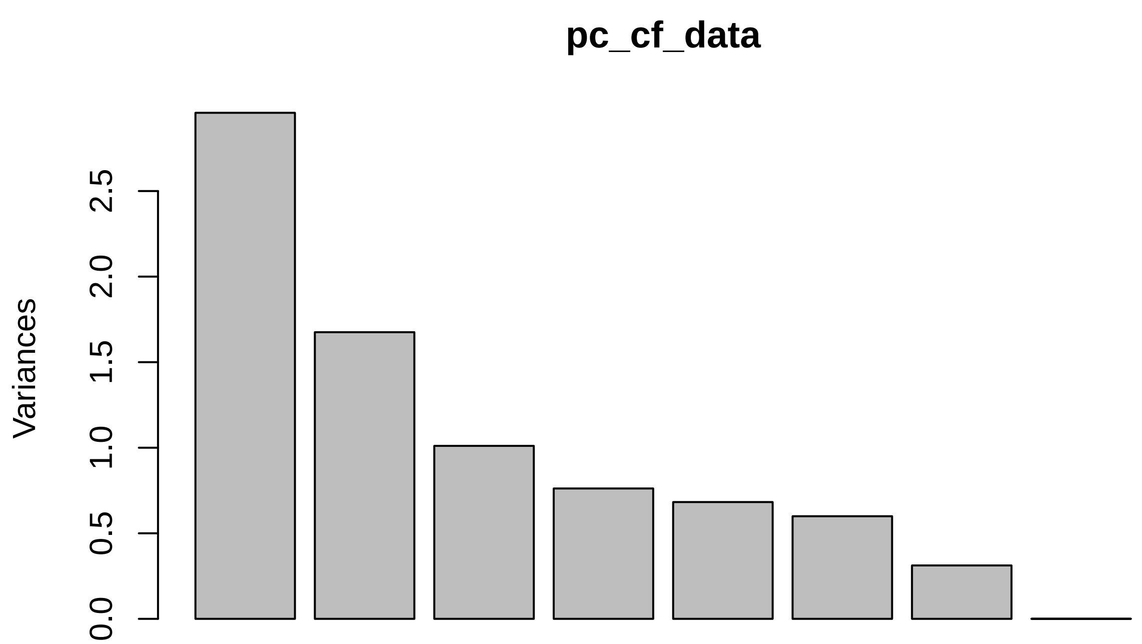
```
# calculate the proportion of exaplained variance (PEV) from the std values
pc_cf_data_var <- pc_cf_data$sdev^2
pc_cf_data_var
```

```
## [1] 2.9573122077 1.6751336598 1.0109269287 0.7618205353 0.6825166481
## [6] 0.5997164209 0.3120470302 0.0005265694
```

```
pc_cf_data_PEV <- pc_cf_data_var / sum(pc_cf_data_var)
pc_cf_data_PEV
```

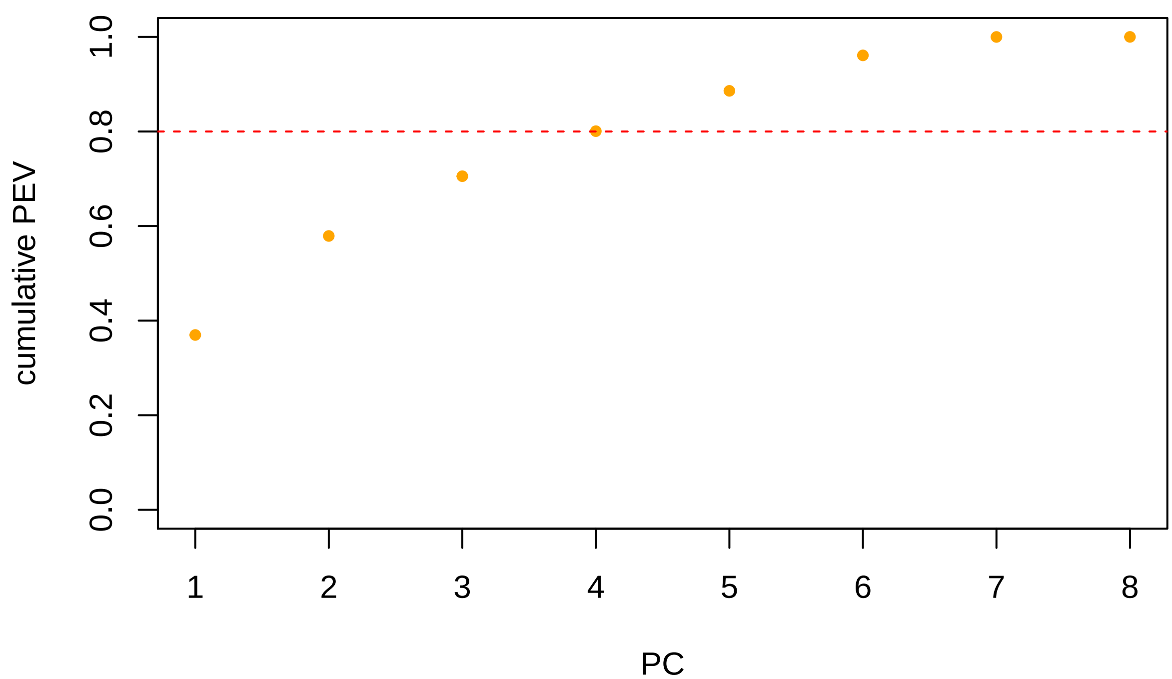
```
## [1] 3.696640e-01 2.093917e-01 1.263659e-01 9.522757e-02 8.531458e-02
## [6] 7.496455e-02 3.900588e-02 6.582117e-05
```

```
# plot the variance per PC
#   note: this can be done using the plot function on the prcomp object
plot(pc_cf_data)
```




```
# plot the cumulative value of PEV for increasing number of additional PCs
#   note: add an 80% threshold line to inform the feature extraction
#     according to the plot the first 3 PCs should be selected
opar <- par()
plot(
  cumsum(pc_cf_data_PEV),
```

```r
  ylim = c(0,1),
  xlab = 'PC',
  ylab = 'cumulative PEV',
  pch = 20,
  col = 'orange'
)
abline(h = 0.8, col = 'red', lty = 'dashed')
```




```r
par(opar)
```

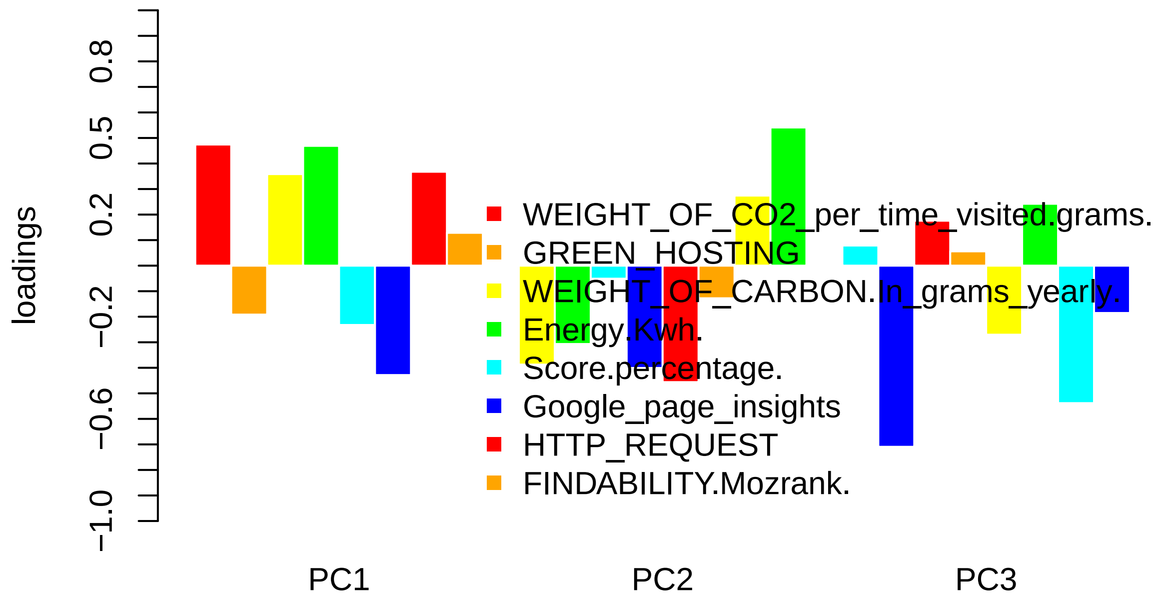
```
## Warning in par(opar): graphical parameter "cin" cannot be set

## Warning in par(opar): graphical parameter "cra" cannot be set

## Warning in par(opar): graphical parameter "csi" cannot be set

## Warning in par(opar): graphical parameter "cxy" cannot be set

## Warning in par(opar): graphical parameter "din" cannot be set

## Warning in par(opar): graphical parameter "page" cannot be set
```

```r
# get and inspect the loadings for each PC
#   note: loadings are reported as a rotation matrix (see lecture)
pc_cf_data_loadings <- pc_cf_data$rotation
pc_cf_data_loadings
```

```
##                                              PC1         PC2         PC3
## WEIGHT_OF_CO2_per_time_visited.grams.  0.4744627 -0.38746110  0.07880132
## GREEN_HOSTING                         -0.1917415 -0.30694039 -0.70953223
## WEIGHT_OF_CARBON.In_grams_yearly.      0.3585778 -0.05105031  0.17611563
## Energy.Kwh.                            0.4691568 -0.40073313  0.05620290
## Score.percentage.                     -0.2323316 -0.45613952 -0.27038104
## Google_page_insights                  -0.4284734 -0.12777996  0.24233669
## HTTP_REQUEST                           0.3678387  0.27439925 -0.53847275
## FINDABILITY.Mozrank.                   0.1286915  0.54109370 -0.18545350
##                                                PC4         PC5        PC6
## WEIGHT_OF_CO2_per_time_visited.grams. -0.079024523 -0.25790963  0.2073403
## GREEN_HOSTING                          0.001864303  0.33171482  0.4926299
## WEIGHT_OF_CARBON.In_grams_yearly.     -0.448130912  0.77700440 -0.1730194
## Energy.Kwh.                           -0.076297527 -0.24849905  0.2215836
## Score.percentage.                     -0.394022017 -0.24591625 -0.6410974
## Google_page_insights                  -0.506770179 -0.04838973  0.2805702
## HTTP_REQUEST                           0.012101806 -0.03958291 -0.2896077
## FINDABILITY.Mozrank.                  -0.612285232 -0.30590377  0.2483803
##                                               PC7          PC8
## WEIGHT_OF_CO2_per_time_visited.grams.  0.03119490  0.708395417
## GREEN_HOSTING                         -0.11142938  0.021164163
## WEIGHT_OF_CARBON.In_grams_yearly.     -0.05747769 -0.001608466
## Energy.Kwh.                            0.04328174 -0.705407983
## Score.percentage.                     -0.19523774 -0.001048991
## Google_page_insights                   0.63517229  0.005891237
## HTTP_REQUEST                           0.64328225  0.006992389
## FINDABILITY.Mozrank.                  -0.35504889 -0.006348386
```

```r
# plot the loadings for the first three PCs as a barplot
#   note: two vectors for colours and labels are created for convenience
#     for details on the other parameters see the help for barplot and legend
opar <- par()
colvector = c('red', 'orange', 'yellow', 'green', 'cyan', 'blue')
labvector = c('PC1', 'PC2', 'PC3')
barplot(
  pc_cf_data_loadings[,c(1:3)],
  beside = T,
  yaxt = 'n',
  names.arg = labvector,
  col = colvector,
  ylim = c(-1,1),
  border = 'white',
  ylab = 'loadings'
)
axis(2, seq(-1,1,0.1))
legend(
  'bottomright',
  bty = 'n',
```

```
  col = colvector,
  pch = 15,
  row.names(pc_cf_data_loadings)
)
```



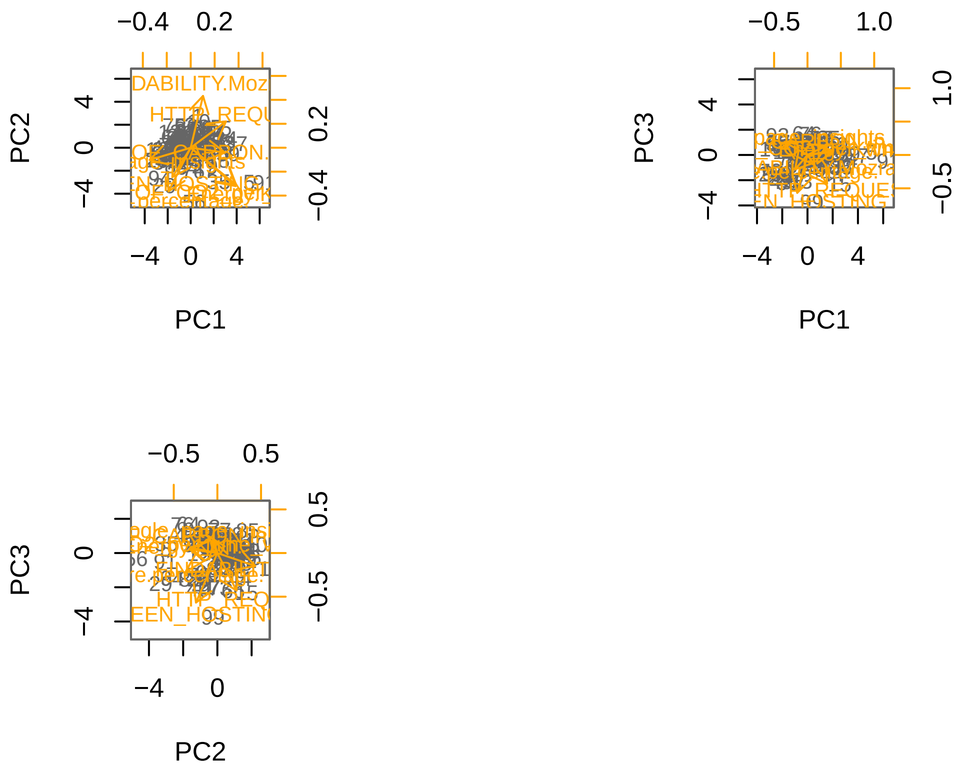

```
par(opar)
```

```
## Warning in par(opar): graphical parameter "cin" cannot be set

## Warning in par(opar): graphical parameter "cra" cannot be set

## Warning in par(opar): graphical parameter "csi" cannot be set

## Warning in par(opar): graphical parameter "cxy" cannot be set

## Warning in par(opar): graphical parameter "din" cannot be set

## Warning in par(opar): graphical parameter "page" cannot be set
```

```
# generate a biplot for each pair of important PCs (and show them on the same page)
#   note: the option choices is used to select the PCs - default is 1:2
opar = par()
par(mfrow = c(2,2))
```

```r
biplot(
  pc_cf_data,
  scale = 0,
  col = c('grey40','orange')
)
biplot(
  pc_cf_data,
  choices = c(1,3),
  scale = 0,
  col = c('grey40','orange')
)
biplot(
  pc_cf_data,
  choices = c(2,3),
  scale = 0,
  col = c('grey40','orange')
)
par(opar)
```

```
## Warning in par(opar): graphical parameter "cin" cannot be set

## Warning in par(opar): graphical parameter "cra" cannot be set

## Warning in par(opar): graphical parameter "csi" cannot be set

## Warning in par(opar): graphical parameter "cxy" cannot be set

## Warning in par(opar): graphical parameter "din" cannot be set

## Warning in par(opar): graphical parameter "page" cannot be set
```




```r
# the space of the first three PCs is better explored interactively...
#   ...using a function from the pca3d package
# first install pca3d
if(require(pca3d) == FALSE){
    install.packages('pca3d')
}
```

```
## Loading required package: pca3d
```

```
## Warning: package 'pca3d' was built under R version 4.0.5
```

```r
# then plot and explore the data by rotating/zoom with the mouse
pca3d::pca3d(pc_cf_data, show.labels = T)
```

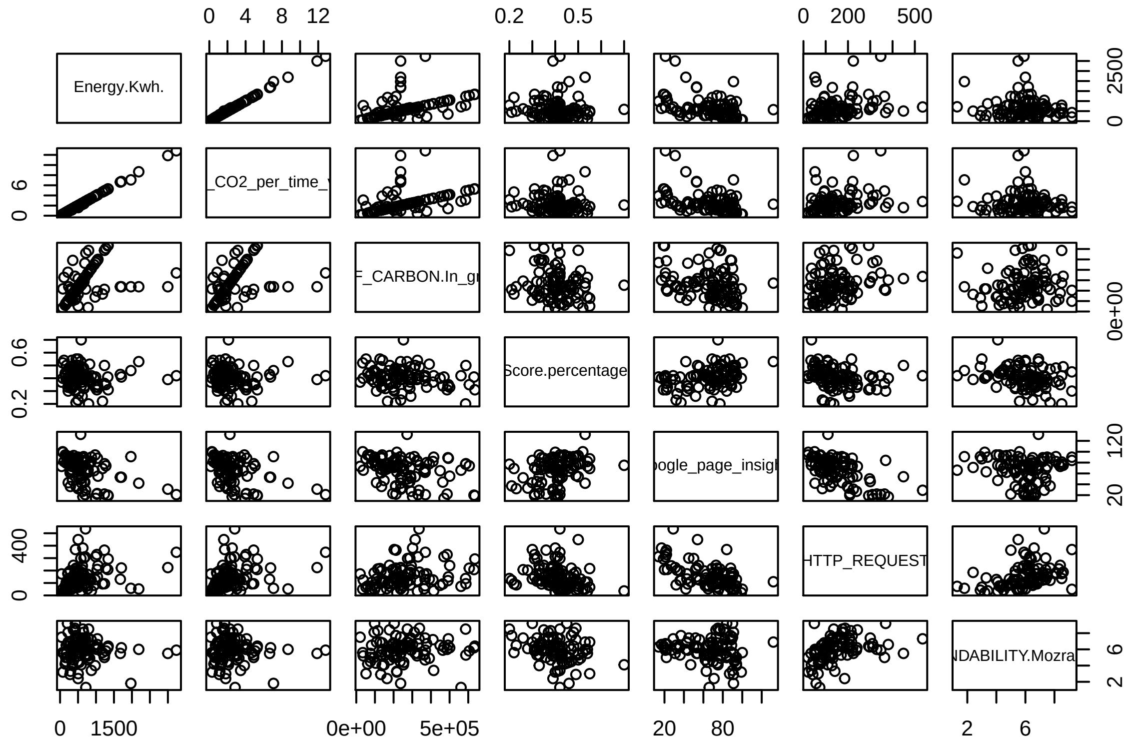
```
## [1] 0.12860729 0.09499318 0.07505453
## Creating new device
```

```r
# and save a snapshot of the view in png format
pca3d::snapshotPCA3d('pc_cf_data_3D.png')
```

From the Principal component analysis we have the line drawn through the 4th PC which means that's how much we have explained variance up to 4 variables.

Using pearson correlation coefficient, Focusing only on the continuous explanatory variables - check their correlations with the Energy. I want to do this only for the continuous variables, so can look to remove the column that is binary from this plot. (This is done so that the pairs plot is legible and that we can run a corr function on the resulting dataframe)

```
cf_data.cont<-subset(cf_data, select=c("Energy.Kwh.", "WEIGHT_OF_CO2_per_time_visited.grams.", "WEIGHT_(
pairs(cf_data.cont)
```




```
cor(cf_data.cont)
```

```
##                                          Energy.Kwh.
## Energy.Kwh.                               1.00000000
## WEIGHT_OF_CO2_per_time_visited.grams.     0.99890976
## WEIGHT_OF_CARBON.In_grams_yearly.         0.41227795
## Score.percentage.                        -0.05473021
## Google_page_insights                     -0.41141346
## HTTP_REQUEST                              0.27177211
## FINDABILITY.Mozrank.                     -0.07952552
##                                          WEIGHT_OF_CO2_per_time_visited.grams.
## Energy.Kwh.                                                         0.99890976
## WEIGHT_OF_CO2_per_time_visited.grams.                               1.00000000
## WEIGHT_OF_CARBON.In_grams_yearly.                                   0.41842680
## Score.percentage.                                                  -0.06608393
## Google_page_insights                                               -0.41886559
## HTTP_REQUEST                                                        0.27162477
## FINDABILITY.Mozrank.                                               -0.06726358
##                                          WEIGHT_OF_CARBON.In_grams_yearly.
```

```
## Energy.Kwh.                                                    0.41227795
## WEIGHT_OF_CO2_per_time_visited.grams.                          0.41842680
## WEIGHT_OF_CARBON.In_grams_yearly.                              1.00000000
## Score.percentage.                                             -0.18137499
## Google_page_insights                                          -0.29344960
## HTTP_REQUEST                                                   0.26412030
## FINDABILITY.Mozrank.                                           0.08457789
##                                       Score.percentage. Google_page_insights
## Energy.Kwh.                                 -0.05473021           -0.4114135
## WEIGHT_OF_CO2_per_time_visited.grams.       -0.06608393           -0.4188656
## WEIGHT_OF_CARBON.In_grams_yearly.           -0.18137499           -0.2934496
## Score.percentage.                            1.00000000            0.3394626
## Google_page_insights                         0.33946265            1.0000000
## HTTP_REQUEST                                -0.24004944           -0.5813457
## FINDABILITY.Mozrank.                        -0.28990679           -0.1064145
##                                       HTTP_REQUEST FINDABILITY.Mozrank.
## Energy.Kwh.                              0.2717721          -0.07952552
## WEIGHT_OF_CO2_per_time_visited.grams.    0.2716248          -0.06726358
## WEIGHT_OF_CARBON.In_grams_yearly.        0.2641203           0.08457789
## Score.percentage.                       -0.2400494          -0.28990679
## Google_page_insights                    -0.5813457          -0.10641454
## HTTP_REQUEST                             1.0000000           0.37787145
## FINDABILITY.Mozrank.                     0.3778715           1.00000000
```

Correlation of the coeficients have been discovered.There do not seem to be any obvious multi collinearity (highly correlated explanatory variables)except the relationship between energy and weight of CO2 per time visited and a few of the plots above point to potential for a linear relationships, therefore at this stage I am not going to explore any transformations.

MACHINE LEARNING (SUPERVISED LEARNING)

Using the continuous explanatory variables decide on a maximal model for Energy and run it.

```r
cf_data.lm<-lm(cf_data$Energy.Kwh.~cf_data$WEIGHT_OF_CO2_per_time_visited.grams.+cf_data$WEIGHT_OF_CARB

summary(cf_data.lm)
```

```
##
## Call:
## lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$WEIGHT_OF_CARBON.In_grams_yearly. + cf_data$Score.percentage. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank.)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.072 -12.571  -4.503   3.106 148.971
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                  -5.532e+00  2.059e+01  -0.269
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  2.538e+02  1.344e+00 188.843
## cf_data$WEIGHT_OF_CARBON.In_grams_yearly.     -1.823e-05  1.889e-05  -0.965
## cf_data$Score.percentage.                      3.984e+01  3.284e+01   1.213
## cf_data$Google_page_insights                   2.588e-01  1.377e-01   1.879
```

```
## cf_data$HTTP_REQUEST                            8.076e-02  3.309e-02   2.440
## cf_data$FINDABILITY.Mozrank.                   -5.057e+00  1.798e+00  -2.813
##                                                Pr(>|t|)
## (Intercept)                                     0.78877
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  < 2e-16 ***
## cf_data$WEIGHT_OF_CARBON.In_grams_yearly.       0.33716
## cf_data$Score.percentage.                       0.22809
## cf_data$Google_page_insights                    0.06341 .
## cf_data$HTTP_REQUEST                            0.01659 *
## cf_data$FINDABILITY.Mozrank.                    0.00601 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.51 on 92 degrees of freedom
## Multiple R-squared:  0.9982, Adjusted R-squared:  0.9981
## F-statistic:  8408 on 6 and 92 DF,  p-value: < 2.2e-16
```

I got a negative intercept and a almost seemingly over fitted model with an Rsquared of 99%. it is possible to start with a model that has interactions, all interactions could be used or a Tree approach can help understand if the relationship between an explanatory variable and the target variable is different based on the value (or range) of the explanatory variable.

So i introduced a step function to get the minimal adequate model.Use a model selection approach to achieve a minimal adequate mode

```
step(cf_data.lm)
```

```
## Start:  AIC=631.94
## cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$WEIGHT_OF_CARBON.In_grams_yearly. + cf_data$Score.percentage. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank.
##
##                                                  Df Sum of Sq      RSS     AIC
## - cf_data$WEIGHT_OF_CARBON.In_grams_yearly.      1       515    51378  630.93
## - cf_data$Score.percentage.                      1       814    51677  631.51
## <none>                                                          50863  631.94
## - cf_data$Google_page_insights                   1      1952    52815  633.66
## - cf_data$HTTP_REQUEST                           1      3293    54156  636.15
## - cf_data$FINDABILITY.Mozrank.                   1      4374    55237  638.10
## - cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  1  19716025 19766888 1220.24
##
## Step:  AIC=630.93
## cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Score.percentage. + cf_data$Google_page_insights +
##     cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank.
##
##                                                  Df Sum of Sq      RSS     AIC
## - cf_data$Score.percentage.                      1       975    52353  630.79
## <none>                                                          51378  630.93
## - cf_data$Google_page_insights                   1      2018    53396  632.75
## - cf_data$HTTP_REQUEST                           1      3097    54475  634.73
## - cf_data$FINDABILITY.Mozrank.                   1      4484    55862  637.22
## - cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  1  22393724 22445102 1230.81
##
```

```
## Step:  AIC=630.79
## cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank.
##
##                                               Df Sum of Sq      RSS     AIC
## <none>                                                         52353  630.79
## - cf_data$Google_page_insights                 1      3229    55583  634.72
## - cf_data$HTTP_REQUEST                         1      3287    55641  634.82
## - cf_data$FINDABILITY.Mozrank.                 1      6061    58415  639.64
## - cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  1  22472695 22525048 1229.17
```

```
##
## Call:
## lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank.)
##
## Coefficients:
##                                  (Intercept)
##                                      7.10635
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.
##                                    253.45616
##                 cf_data$Google_page_insights
##                                      0.31753
##                         cf_data$HTTP_REQUEST
##                                      0.08029
##                 cf_data$FINDABILITY.Mozrank.
##                                     -5.74259
```

My minimal adequte model has been achieved.Once I have the minimal adequate model, explain its findings and test its residuals

```
mam.lm<-lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. + cf_data$Googl
    cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank.)

summary(mam.lm)
```

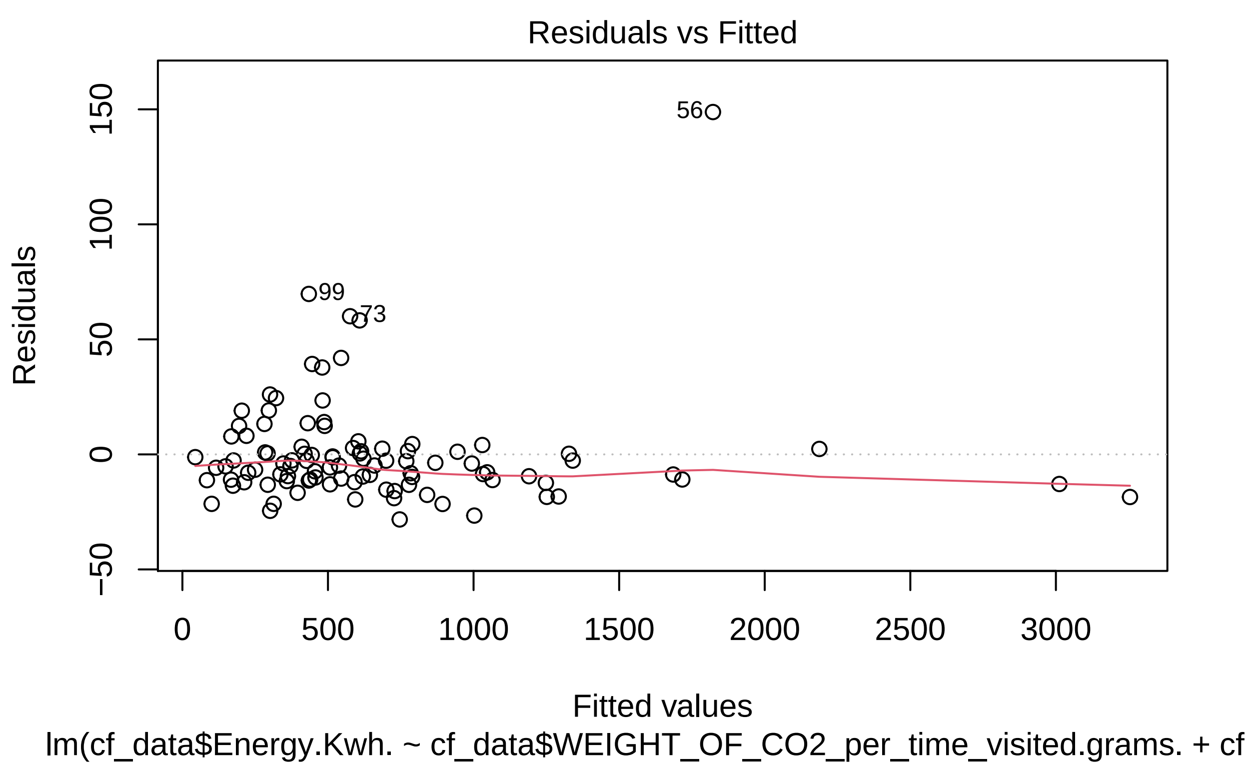
```
##
## Call:
## lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank.)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.275 -11.286  -4.864   2.427 148.823
##
## Coefficients:
##                                              Estimate Std. Error t value
## (Intercept)                                   7.10635   15.42750   0.461
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams. 253.45616    1.26178 200.872
## cf_data$Google_page_insights                  0.31753    0.13187   2.408
## cf_data$HTTP_REQUEST                          0.08029    0.03305   2.429
## cf_data$FINDABILITY.Mozrank.                 -5.74259    1.74073  -3.299
##                                             Pr(>|t|)
```

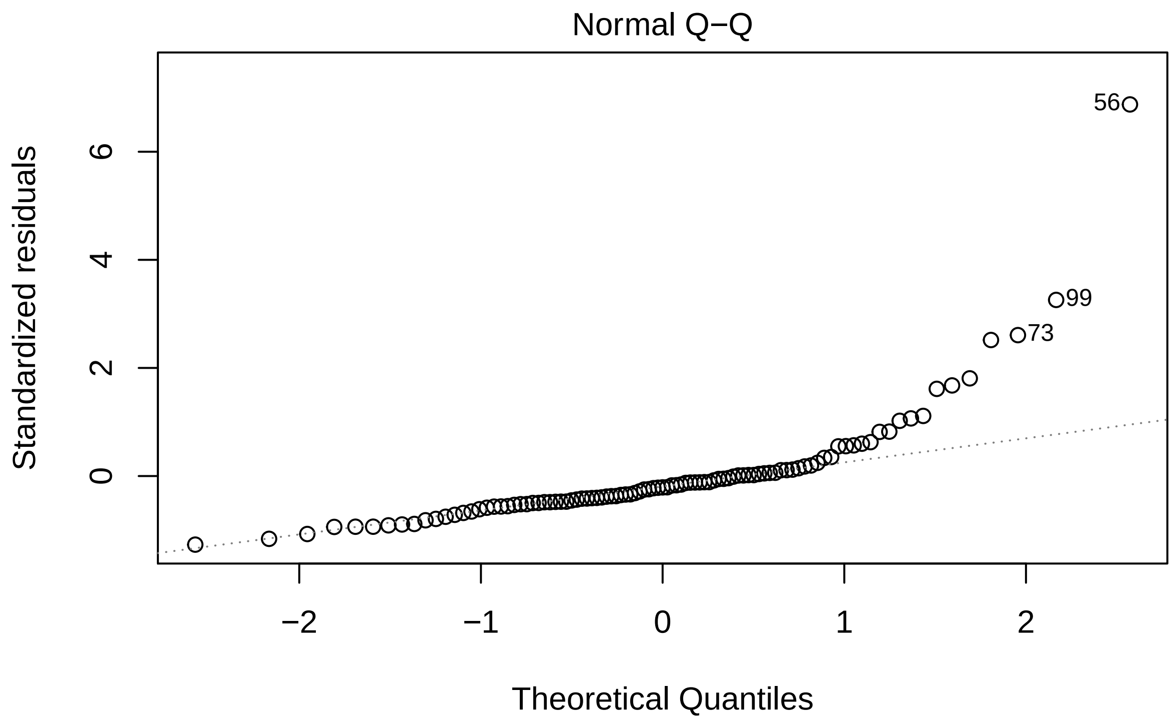
```
## (Intercept)                                      0.64613
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  < 2e-16 ***
## cf_data$Google_page_insights                     0.01799 *
## cf_data$HTTP_REQUEST                             0.01702 *
## cf_data$FINDABILITY.Mozrank.                     0.00137 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.6 on 94 degrees of freedom
## Multiple R-squared:  0.9981, Adjusted R-squared:  0.998
## F-statistic: 1.252e+04 on 4 and 94 DF,  p-value: < 2.2e-16
```

This model has acceptable goodness of fit, all the coefficients are significant (so there is no need to simplyfy further), r^2 is too high and the F statistic is significant.

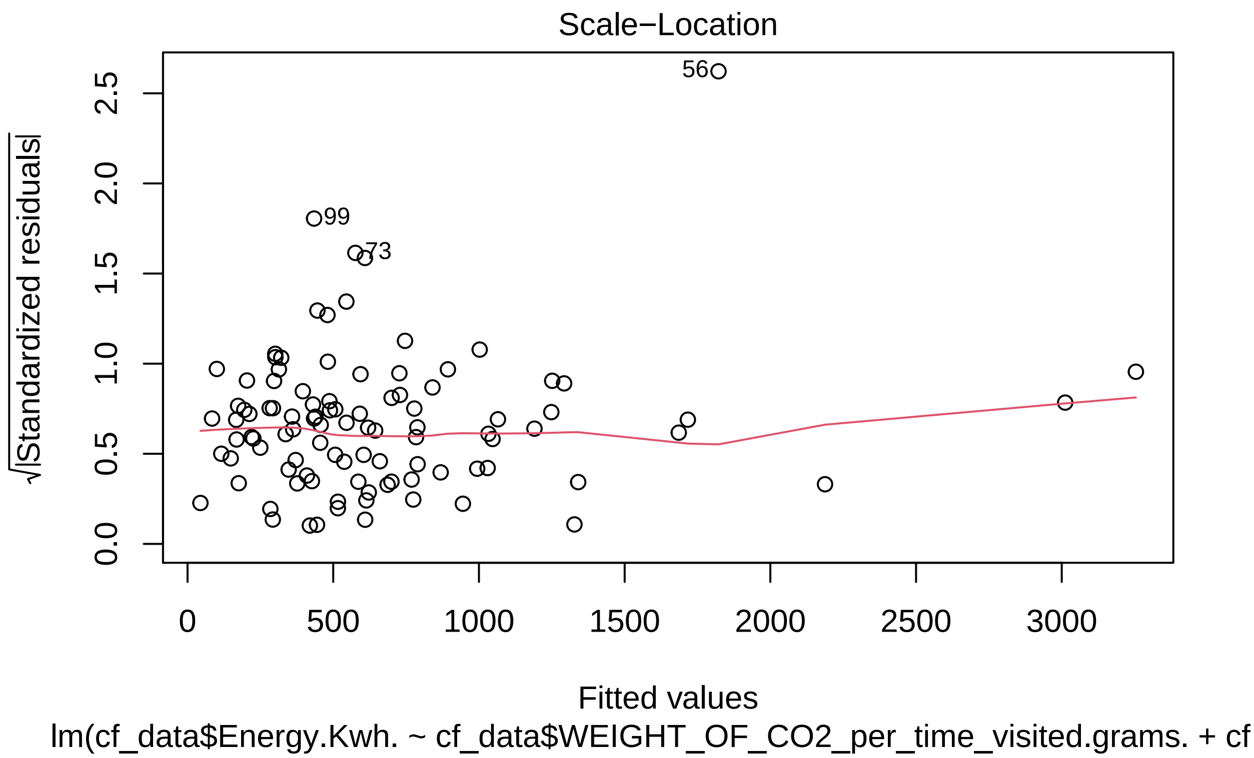
Next the residuals should be scrutinised:

```
plot(mam.lm)
```

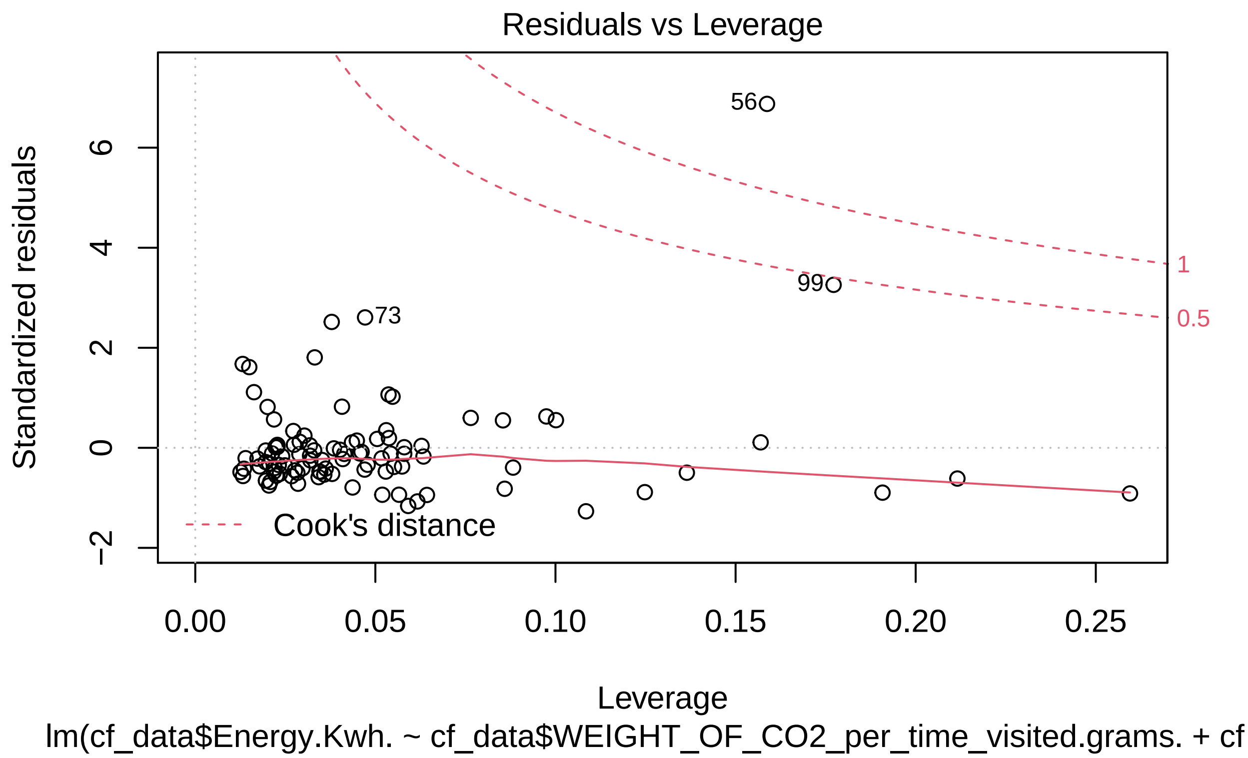




lm(cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. + cf



Scale−Location
56
99
73
0.0
0.5
1.0
1.5
2.0
2.5
√|Standardized residuals|
0
500
1000
1500
2000
2500
3000
Fitted values
lm(cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. + cf



In this case the residuals look ok, the variance is quite steady in the first plot - considering the data size. QQ plot also looks aligned.

Now i want to model the relationship between the energy and the explanatory variables (including the ones that are not continuous).

```
model.all.lm<-lm(cf_data$Energy.Kwh.~cf_data$WEIGHT_OF_CO2_per_time_visited.grams.+cf_data$WEIGHT_OF_CA

summary(model.all.lm)
```

```
##
## Call:
## lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$WEIGHT_OF_CARBON.In_grams_yearly. + cf_data$Score.percentage. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank. +
##     cf_data$GREEN_HOSTING)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.340  -5.330  -0.914   2.727 117.132
##
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                   1.739e+00  1.581e+01   0.110
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  2.544e+02  1.033e+00 246.344
## cf_data$WEIGHT_OF_CARBON.In_grams_yearly.     -7.655e-06  1.454e-05  -0.526
```

```
## cf_data$Score.percentage.                         -9.299e+00  2.589e+01  -0.359
## cf_data$Google_page_insights                       1.827e-01  1.060e-01   1.724
## cf_data$HTTP_REQUEST                               5.494e-02  2.556e-02   2.149
## cf_data$FINDABILITY.Mozrank.                      -3.210e+00  1.397e+00  -2.298
## cf_data$GREEN_HOSTING                              4.565e+01  5.636e+00   8.099
##                                                   Pr(>|t|)
## (Intercept)                                         0.9126
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  < 2e-16 ***
## cf_data$WEIGHT_OF_CARBON.In_grams_yearly.           0.5999
## cf_data$Score.percentage.                           0.7203
## cf_data$Google_page_insights                        0.0881 .
## cf_data$HTTP_REQUEST                                0.0343 *
## cf_data$FINDABILITY.Mozrank.                        0.0239 *
## cf_data$GREEN_HOSTING                             2.39e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.02 on 91 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
## F-statistic: 1.228e+04 on 7 and 91 DF,  p-value: < 2.2e-16
```

The $r^2 $is looking same but lets see what a step process would acheive in terms of simplifying the model:

```r
step(model.all.lm)
```

```
## Start:  AIC=580.2
## cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$WEIGHT_OF_CARBON.In_grams_yearly. + cf_data$Score.percentage. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank. +
##     cf_data$GREEN_HOSTING
##
##                                                   Df Sum of Sq      RSS     AIC
## - cf_data$Score.percentage.                        1        42    29600  578.34
## - cf_data$WEIGHT_OF_CARBON.In_grams_yearly.        1        90    29648  578.50
## <none>                                                            29558  580.20
## - cf_data$Google_page_insights                     1       965    30523  581.38
## - cf_data$HTTP_REQUEST                             1      1500    31058  583.10
## - cf_data$FINDABILITY.Mozrank.                     1      1715    31273  583.78
## - cf_data$GREEN_HOSTING                            1     21305    50863  631.94
## - cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  1  19711502 19741060 1222.11
##
## Step:  AIC=578.34
## cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$WEIGHT_OF_CARBON.In_grams_yearly. + cf_data$Google_page_insights +
##     cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank. + cf_data$GREEN_HOSTING
##
##                                                   Df Sum of Sq      RSS     AIC
## - cf_data$WEIGHT_OF_CARBON.In_grams_yearly.        1        80    29680  576.61
## <none>                                                            29600  578.34
## - cf_data$Google_page_insights                     1       926    30526  579.39
## - cf_data$HTTP_REQUEST                             1      1487    31087  581.19
## - cf_data$FINDABILITY.Mozrank.                     1      1677    31278  581.80
## - cf_data$GREEN_HOSTING                            1     22077    51677  631.51
```

```
## - cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  1  19911421 19941021 1221.10
##
## Step:  AIC=576.61
## cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank. +
##     cf_data$GREEN_HOSTING
##
##                                                  Df Sum of Sq      RSS     AIC
## <none>                                                          29680  576.61
## - cf_data$Google_page_insights                    1       955    30635  577.74
## - cf_data$HTTP_REQUEST                            1      1436    31116  579.29
## - cf_data$FINDABILITY.Mozrank.                    1      1708    31389  580.15
## - cf_data$GREEN_HOSTING                           1     22673    52353  630.79
## - cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  1  22416995 22446675 1230.82
```

```
##
## Call:
## lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank. +
##     cf_data$GREEN_HOSTING)
##
## Coefficients:
##                                  (Intercept)
##                                     -3.19522
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.
##                                    254.18927
##                 cf_data$Google_page_insights
##                                      0.17514
##                         cf_data$HTTP_REQUEST
##                                      0.05349
##                 cf_data$FINDABILITY.Mozrank.
##                                     -3.13167
##                        cf_data$GREEN_HOSTING
##                                     45.48271
```

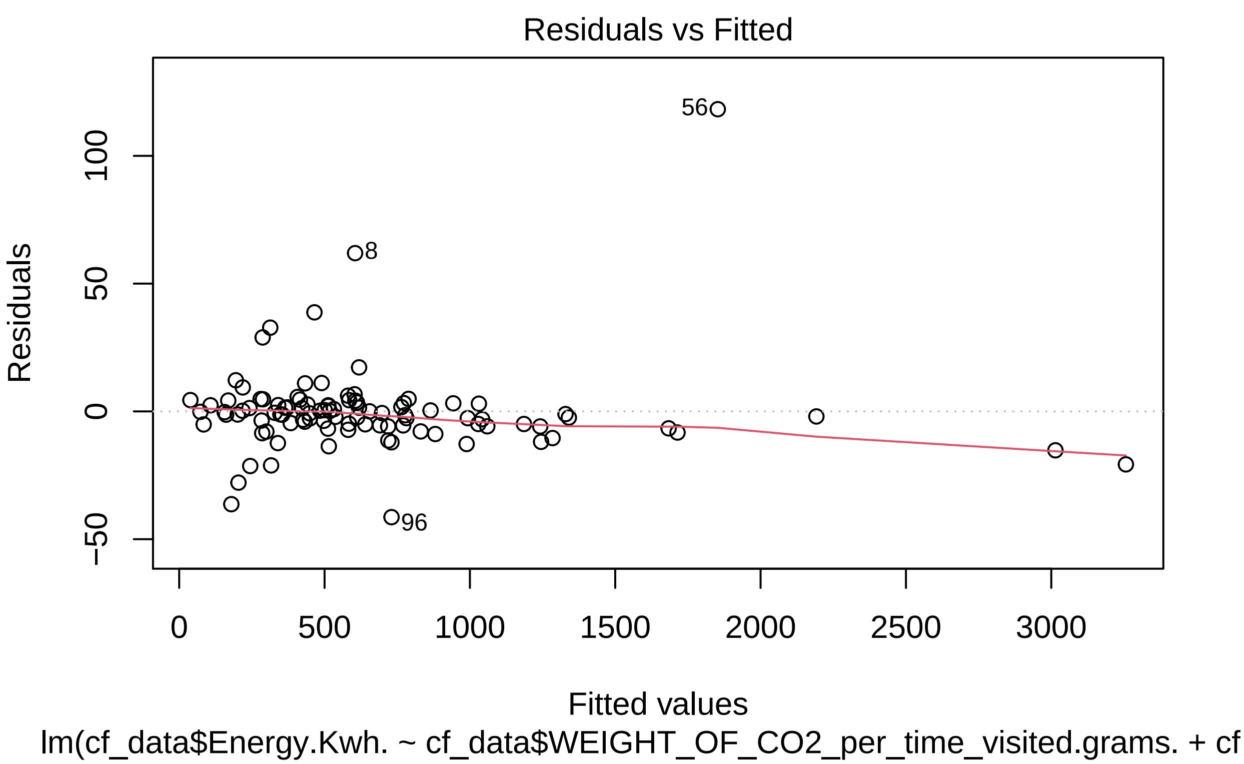
It is evident Greenhosting has an effect on this model so i would explore it further.The binary variable I added as part of the explanatory variables does add much and this is confirmed as the step process proposes a model that does include it as an explanatory variable.

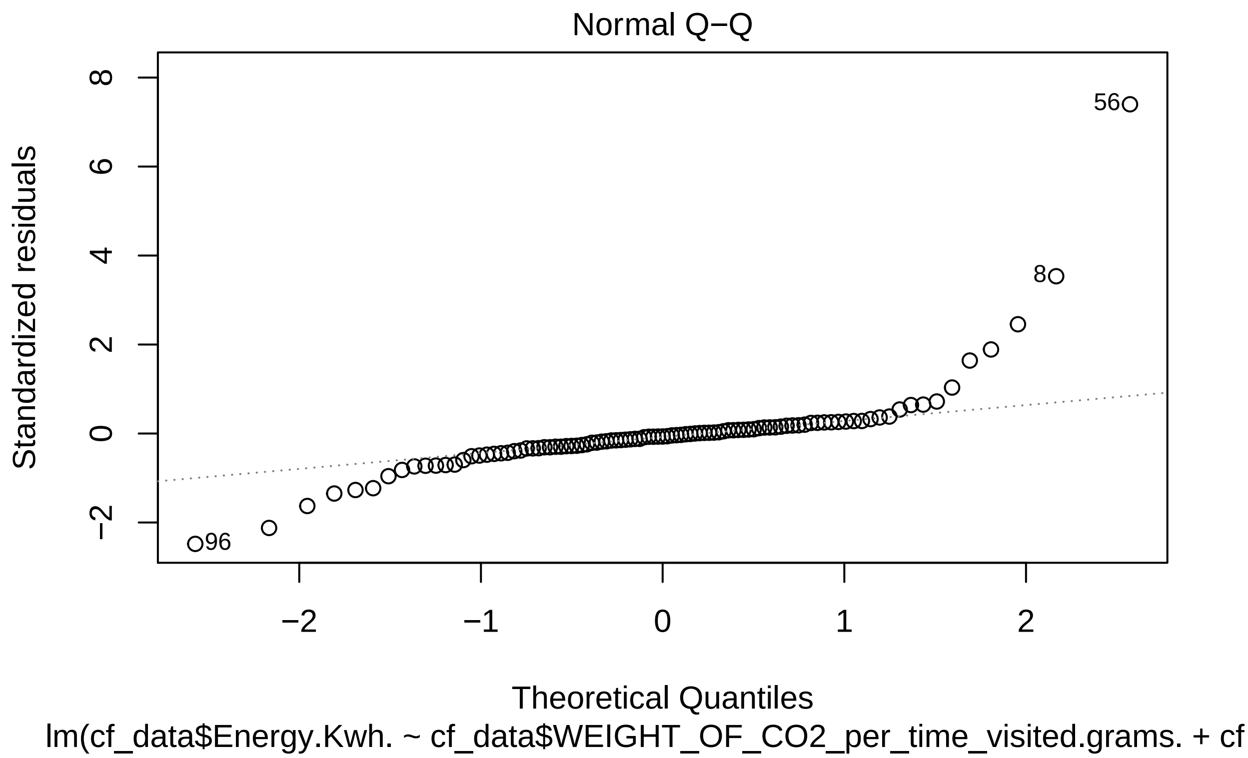
```r
all.mam.lm<-lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. + cf_data$(

summary(all.mam.lm)
```

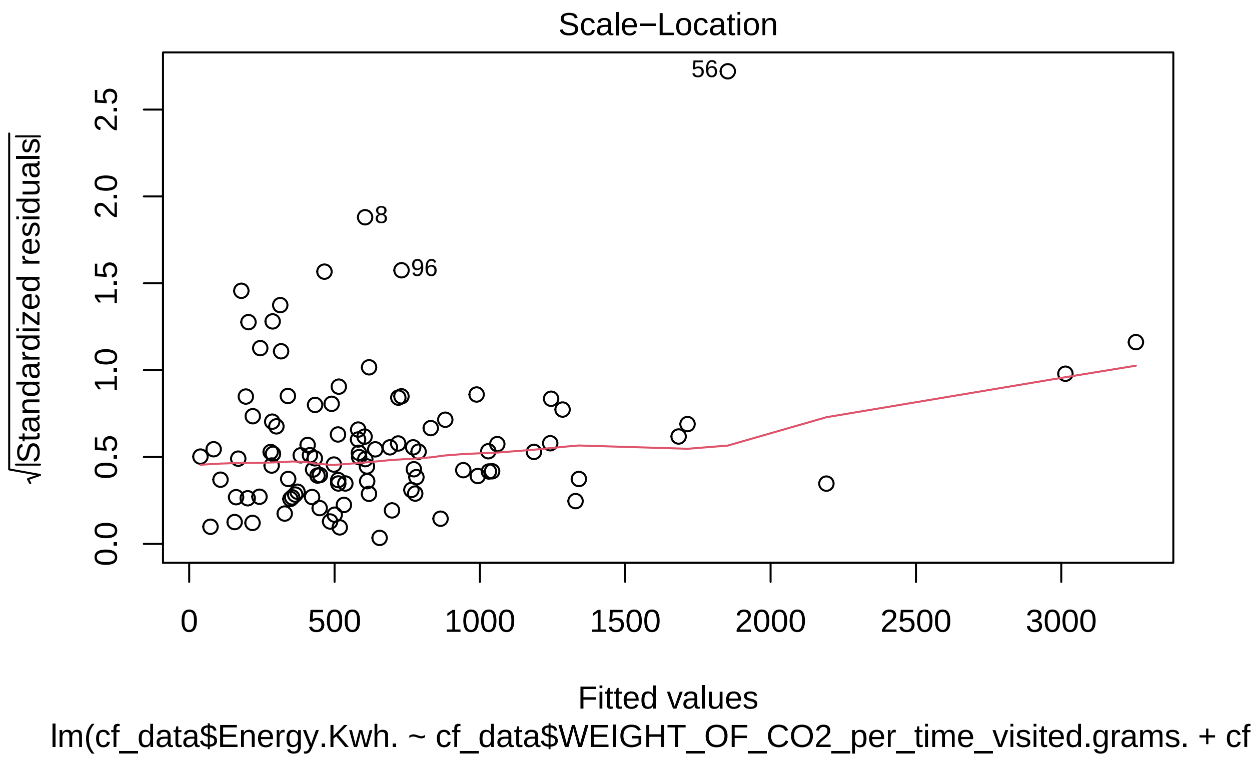
```
##
## Call:
## lm(formula = cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank. +
##     cf_data$GREEN_HOSTING)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.374  -5.650  -1.175   2.866 118.245
##
```

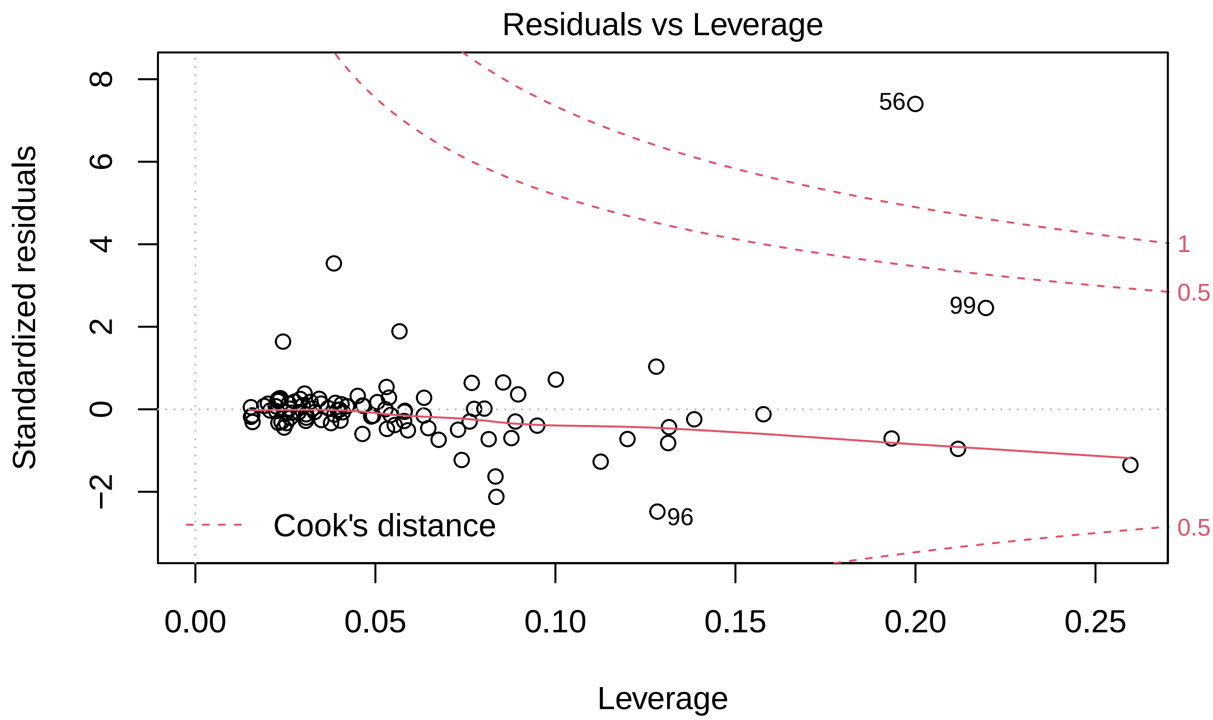
```
## Coefficients:
##                                                Estimate Std. Error t value
## (Intercept)                                    -3.19522   11.74211  -0.272
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams. 254.18927    0.95910 265.030
## cf_data$Google_page_insights                    0.17514    0.10124   1.730
## cf_data$HTTP_REQUEST                            0.05349    0.02522   2.121
## cf_data$FINDABILITY.Mozrank.                   -3.13167    1.35362  -2.314
## cf_data$GREEN_HOSTING                          45.48271    5.39617   8.429
##                                               Pr(>|t|)
## (Intercept)                                     0.7861
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  < 2e-16 ***
## cf_data$Google_page_insights                    0.0870 .
## cf_data$HTTP_REQUEST                            0.0366 *
## cf_data$FINDABILITY.Mozrank.                    0.0229 *
## cf_data$GREEN_HOSTING                         4.28e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.86 on 93 degrees of freedom
## Multiple R-squared:  0.9989, Adjusted R-squared:  0.9989
## F-statistic: 1.749e+04 on 5 and 93 DF,  p-value: < 2.2e-16
```

```r
plot(all.mam.lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. + cf
56
8
96
−2
0
2
4
6
8
−2
−1
0
1
2


Scale−Location
56
8
96
$\sqrt{|\text{Standardized residuals}|}$
0.0
0.5
1.0
1.5
2.0
2.5
0
500
1000
1500
2000
2500
3000
Fitted values
lm(cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. + cf



lm(cf_data$Energy.Kwh. ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. + cf

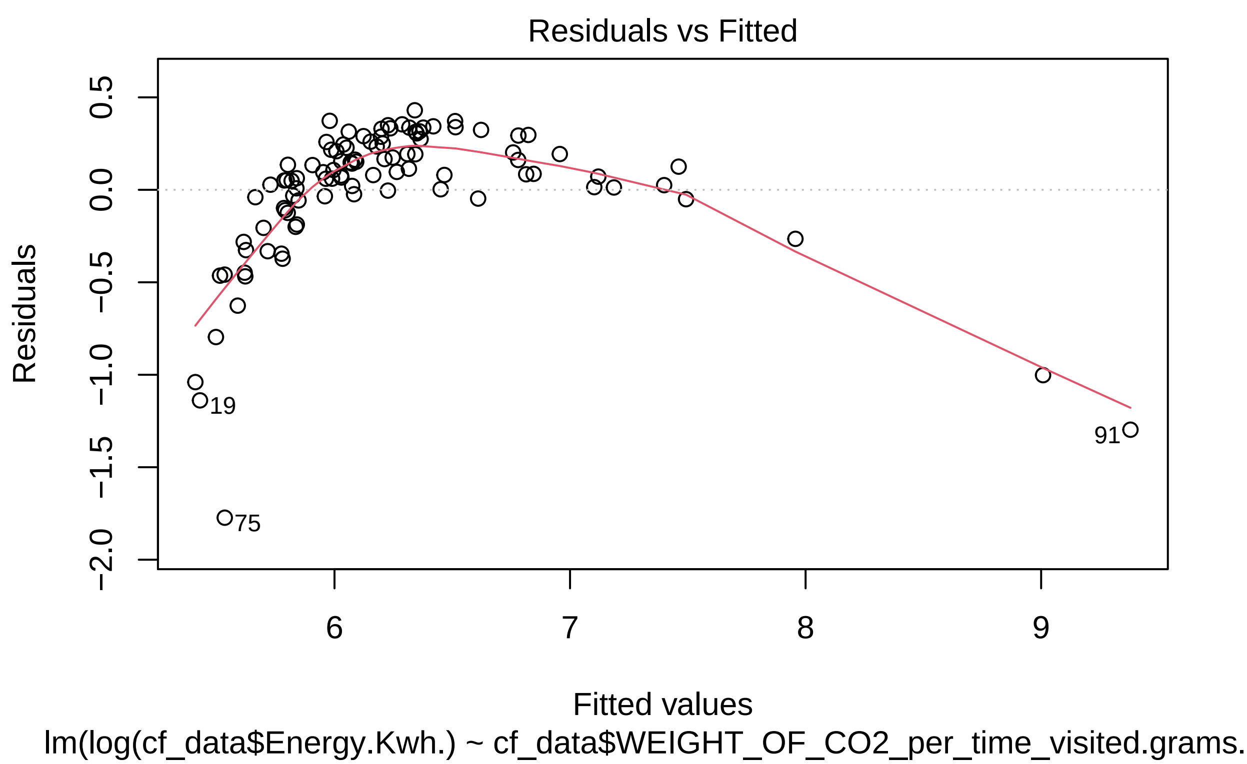
Now i have to optimise the model to reduce the chances of error. I would use the log transformation method to do this.

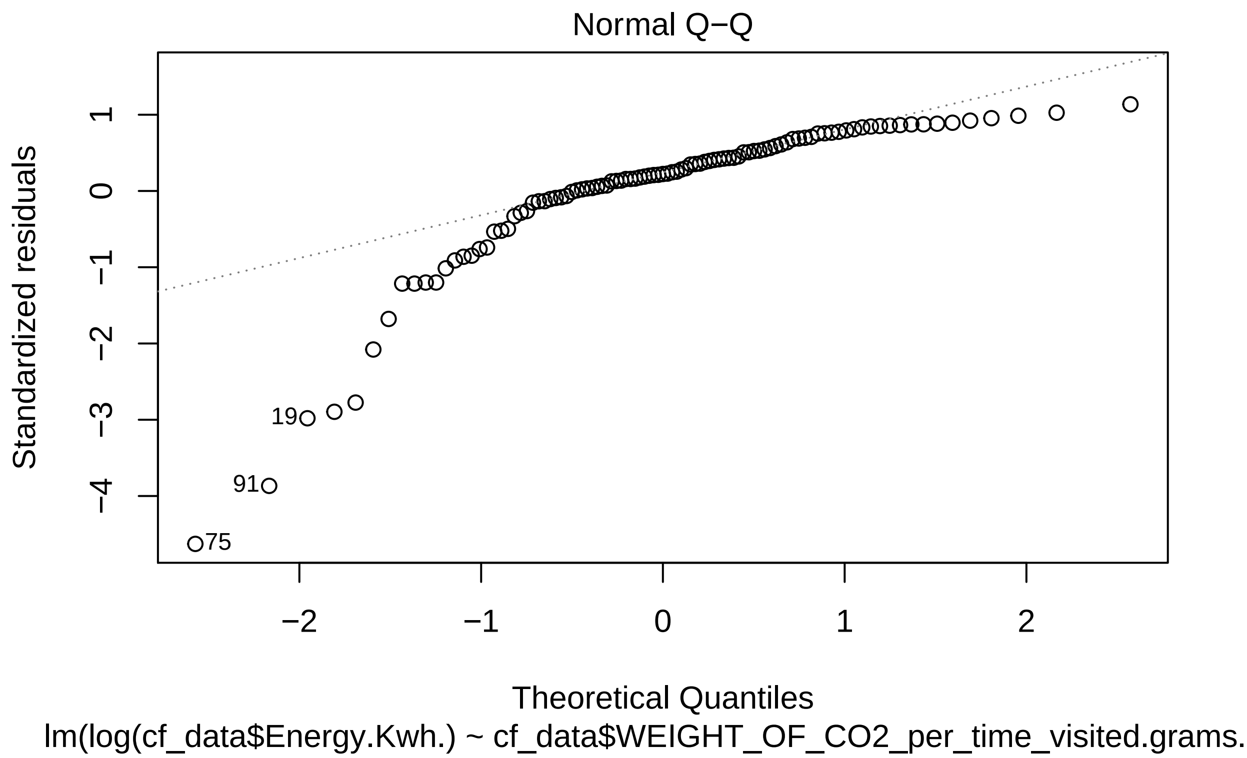
```
optimised_mam.lm<-lm(formula = log(cf_data$Energy.Kwh.) ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams.

summary(optimised_mam.lm)
```

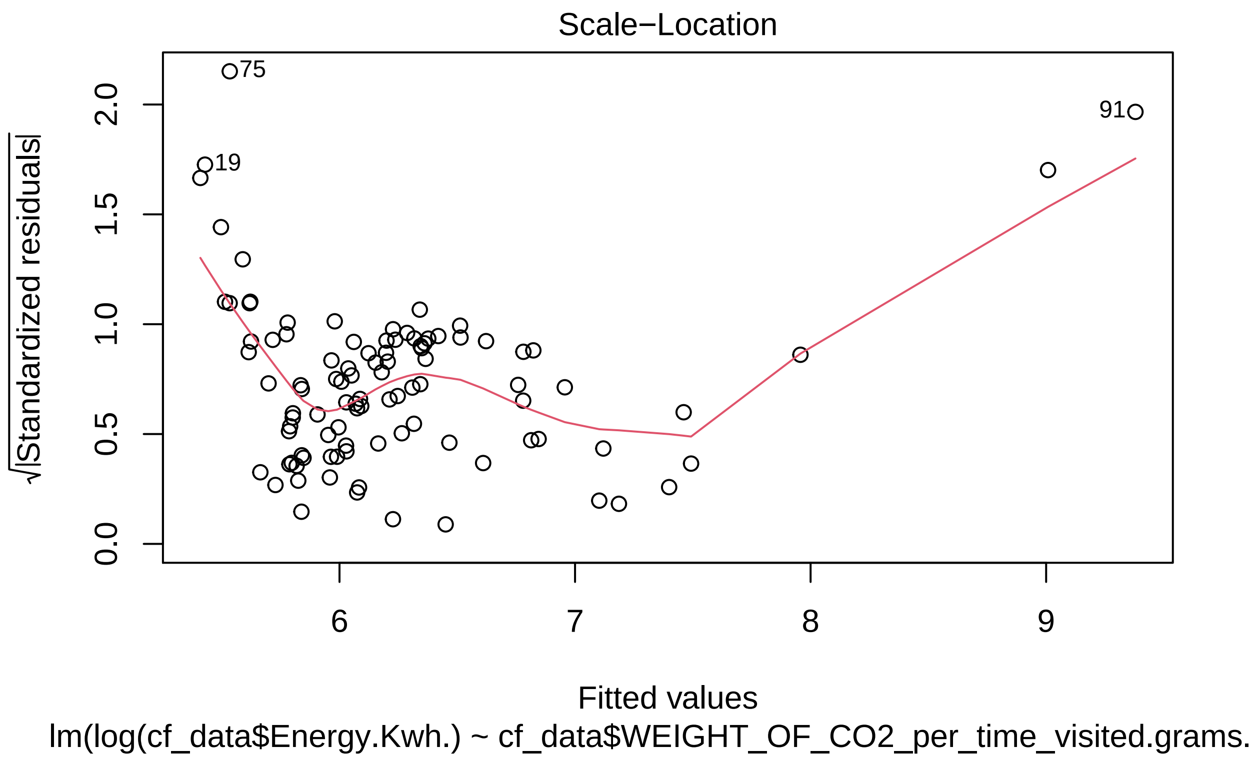
```
##
## Call:
## lm(formula = log(cf_data$Energy.Kwh.) ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams. +
##     cf_data$Google_page_insights + cf_data$HTTP_REQUEST + cf_data$FINDABILITY.Mozrank. +
##     cf_data$GREEN_HOSTING)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.77286 -0.04904  0.08396  0.23954  0.42997
##
## Coefficients:
##                                                 Estimate Std. Error t value
## (Intercept)                                    5.4872534  0.2562362  21.415
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  0.2877573  0.0209294  13.749
## cf_data$Google_page_insights                  -0.0019596  0.0022093  -0.887
## cf_data$HTTP_REQUEST                           0.0006451  0.0005503   1.172
## cf_data$FINDABILITY.Mozrank.                   0.0039111  0.0295387   0.132
## cf_data$GREEN_HOSTING                          0.0739018  0.1177552   0.628
##                                               Pr(>|t|)
```

```
## (Intercept)                                     <2e-16 ***
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.   <2e-16 ***
## cf_data$Google_page_insights                     0.377
## cf_data$HTTP_REQUEST                             0.244
## cf_data$FINDABILITY.Mozrank.                     0.895
## cf_data$GREEN_HOSTING                            0.532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3898 on 93 degrees of freedom
## Multiple R-squared:  0.7429, Adjusted R-squared:  0.7291
## F-statistic: 53.75 on 5 and 93 DF,  p-value: < 2.2e-16
```

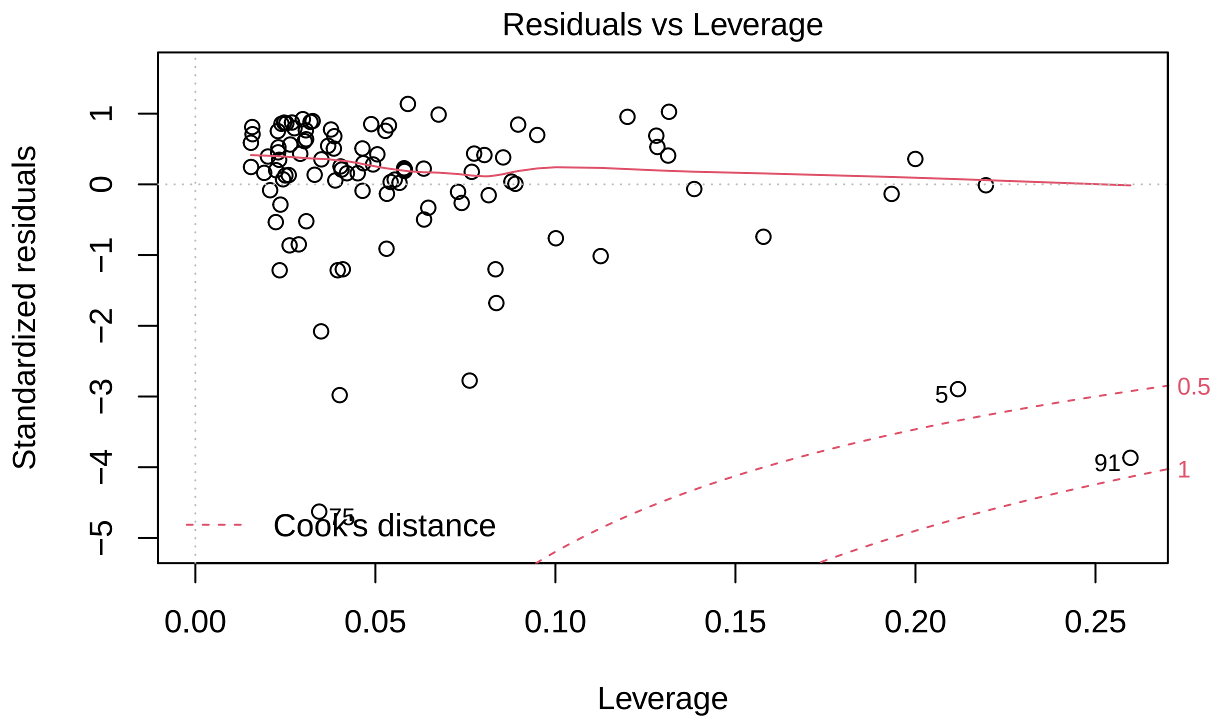
```r
plot(optimised_mam.lm)
```




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(cf_data$Energy.Kwh.) ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams.
1
0
−1
−2
−3
−4
−2
−1
0
1
2
19
91
75



Scale−Location
75
91
19
2.0
1.5
1.0
0.5
0.0
√|Standardized residuals|
6
7
8
9
Fitted values
lm(log(cf_data$Energy.Kwh.) ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams.



lm(log(cf_data$Energy.Kwh.) ~ cf_data$WEIGHT_OF_CO2_per_time_visited.grams.

Now it is evident from my result this model is very significant owing from the value of its Adjusted Rsquared which is 73% and its F-statistic.

Now i move on to calculate my Confidence Interval and Sigma (residual standard error)

```
#Calculating the sigma

sigma(optimised_mam.lm)/mean(cf_data$Energy.Kwh.)
```

```
## [1] 0.0005894417
```

```
#calculating the confidence interval
confint(optimised_mam.lm)
```

```
##                                                    2.5 %      97.5 %
## (Intercept)                                  4.978419016 5.996087724
## cf_data$WEIGHT_OF_CO2_per_time_visited.grams.  0.246195696 0.329318894
## cf_data$Google_page_insights                 -0.006346765 0.002427558
## cf_data$HTTP_REQUEST                         -0.000447737 0.001738000
## cf_data$FINDABILITY.Mozrank.                 -0.054746855 0.062569088
## cf_data$GREEN_HOSTING                        -0.159936662 0.307740277
```

I am 97.5% confident my mean is between 4.978 and 5.996. I also do have a good value of sigma which is 0.00059.

From the values gotten from my model, this model can be used to predict the energy produced by other variables that make up the carbon footprint of companies.