Zeppelin

Performance Analysis and Code (There are a total of 195 lines of code)

FINISHED

Imported the packages

FINISHED

```
import org.apache.spark.sql.functions._
import org.joda.time.format.DateTimeFormat
import org.apache.spark.ml.regression.LinearRegression
import org.apache.spark.ml.regression.LinearRegression
import org.apache.spark.mllib.util.MLUtils
```

```
import org.apache.spark.sql.functions._
import org.joda.time.format.DateTimeFormat
import org.apache.spark.ml.regression.LinearRegression
import org.apache.spark.ml.regression.LinearRegression
import org.apache.spark.mllib.util.MLUtils
```

Adjusted the path to the location of the data

FINISHED

```
// Load data - adjust the path to the location of your data
val inputPath =  "/Users/joannariascos/Desktop/algorithm/aarhus_parking.csv"
val parkingdata = sqlContext.read
        .format("com.databricks.spark.csv")
        .option("header", "true") // Use first line of all files as header
        .option("delimiter", ",")
        .option("inferSchema", "true") // Automatically infer data types
        .load(inputPath)
        parkingdata.registerTempTable("parkingdata")
```

```
inputPath: String = /Users/joannariascos/Desktop/algorithm/aarhus_parking.csv
parkingdata: org.apache.spark.sql.DataFrame = [vehiclecount: int, totalspaces: int ... 2 mo
re fields]
warning: there was one deprecation warning; re-run with -deprecation for details
```

Created the RDD pairs

FINISHED

```
//To read the file
val csv = sc.textFile("/Users/joannariascos/Desktop/algorithm/aarhus_parking.csv");
//To find the headers
val header = csv.first;
```

```
//To remove the header
val data = csv.filter(_(0) != header(0));
//To create a RDD of (label, features) pairs
val parsedData = data.map { line =>
    val parts = line.split(',')
    LabeledPoint(parts(0).toDouble, Vectors.dense(parts(1).split(' ').map(_.toDouble)
    }.cache()
```

```
csv: org.apache.spark.rdd.RDD[String] = /Users/joannariascos/Desktop/algorithm/aarhus_parki
ng.csv MapPartitionsRDD[49] at textFile at <console>:42
header: String = vehiclecount,totalspaces,garagecode,ozone
data: org.apache.spark.rdd.RDD[String] = MapPartitionsRDD[50] at filter at <console>:45
parsedData: org.apache.spark.rdd.RDD[org.apache.spark.mllib.regression.LabeledPoint] = MapP
artitionsRDD[51] at map at <console>:47
```

Loaded the parking dataset with spark

FINISHED

```
%spark.r
aarhus_parking <- read.csv("/Users/joannariascos/Desktop/algorithm/aarhus_parking.csv'
head(aarhus_parking)
```

```
vehiclecount totalspaces   garagecode ozone
1            0          65    NORREPORT   101
2            0         512  SKOLEBAKKEN   106
3          869        1240  SCANDCENTER   107
4           22         953       BRUUNS   103
5          124         130 BUSGADEHUSET   105
6          106         400      MAGASIN   106
```

Fitted the model and ran a multiple regression analysis

FINISHED

```
%r
model = lm(ozone~vehiclecount+totalspaces+garagecode, data = aarhus_parking)
```

Created the anova table

FINISHED

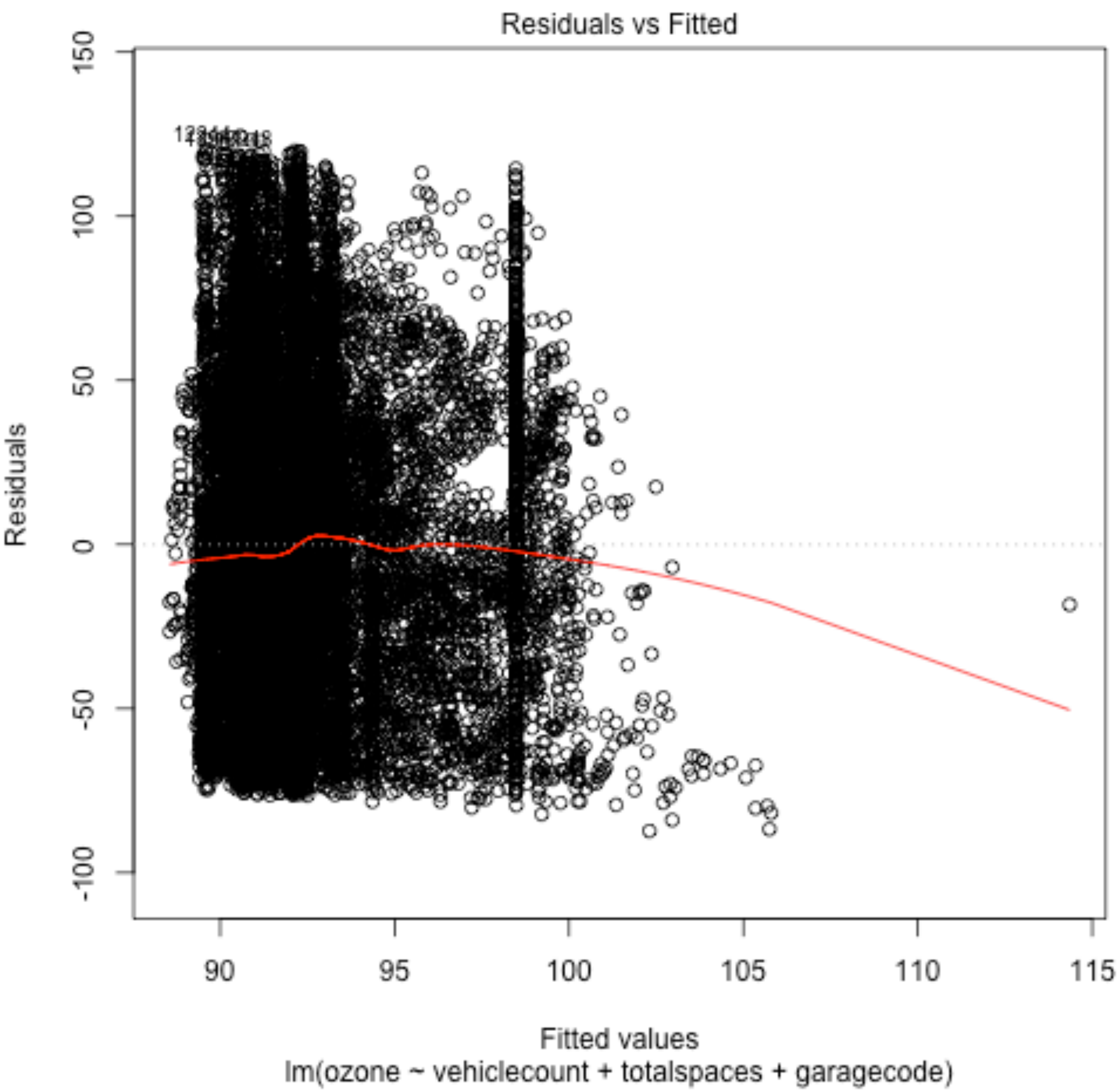
```
%r
modeltwo = lm(ozone~totalspaces, data = aarhus_parking)
anova(model,modeltwo)
```

```
Analysis of Variance Table
```

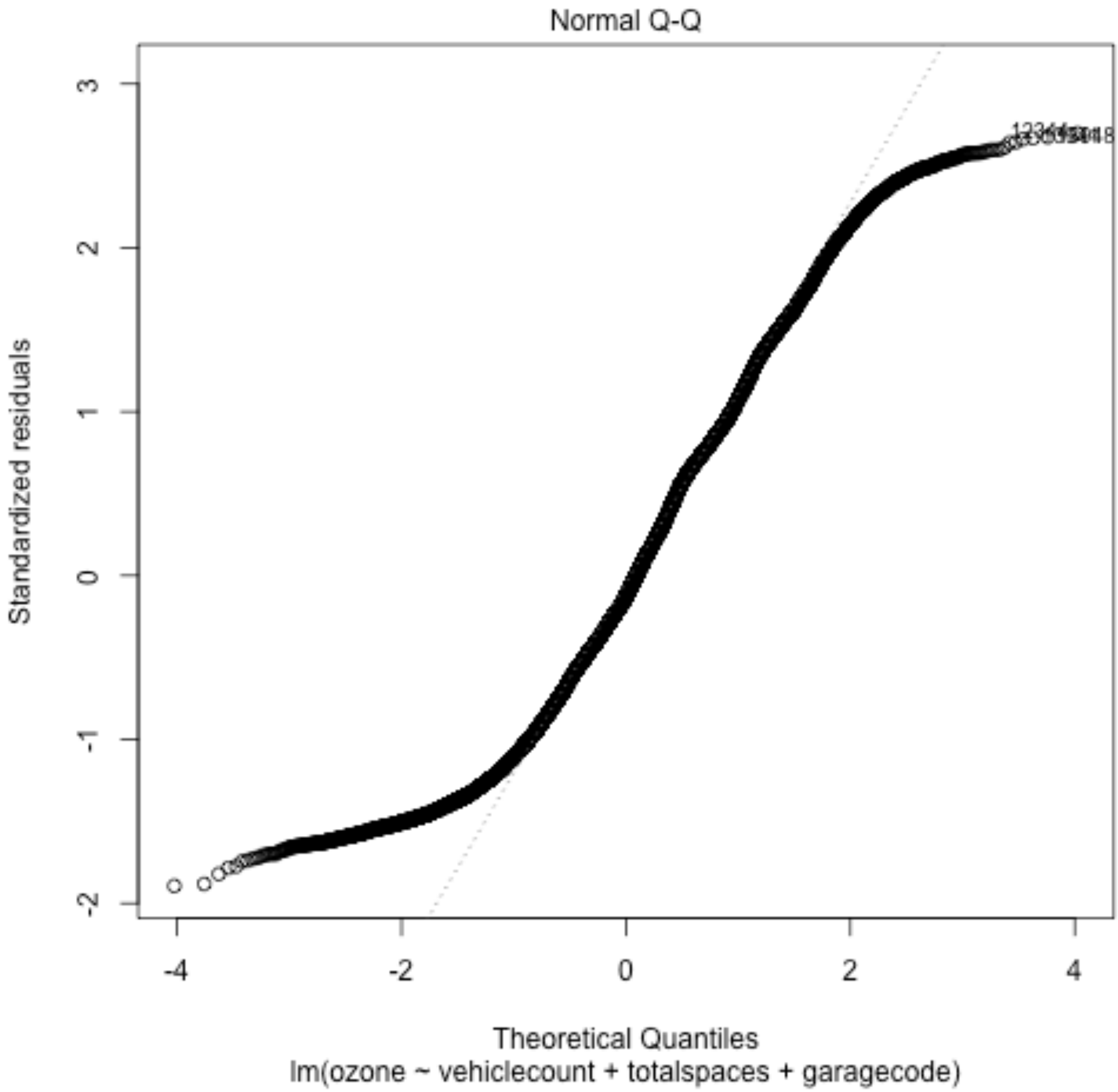
Plotted a residuals vs fitted graph

FINISHED

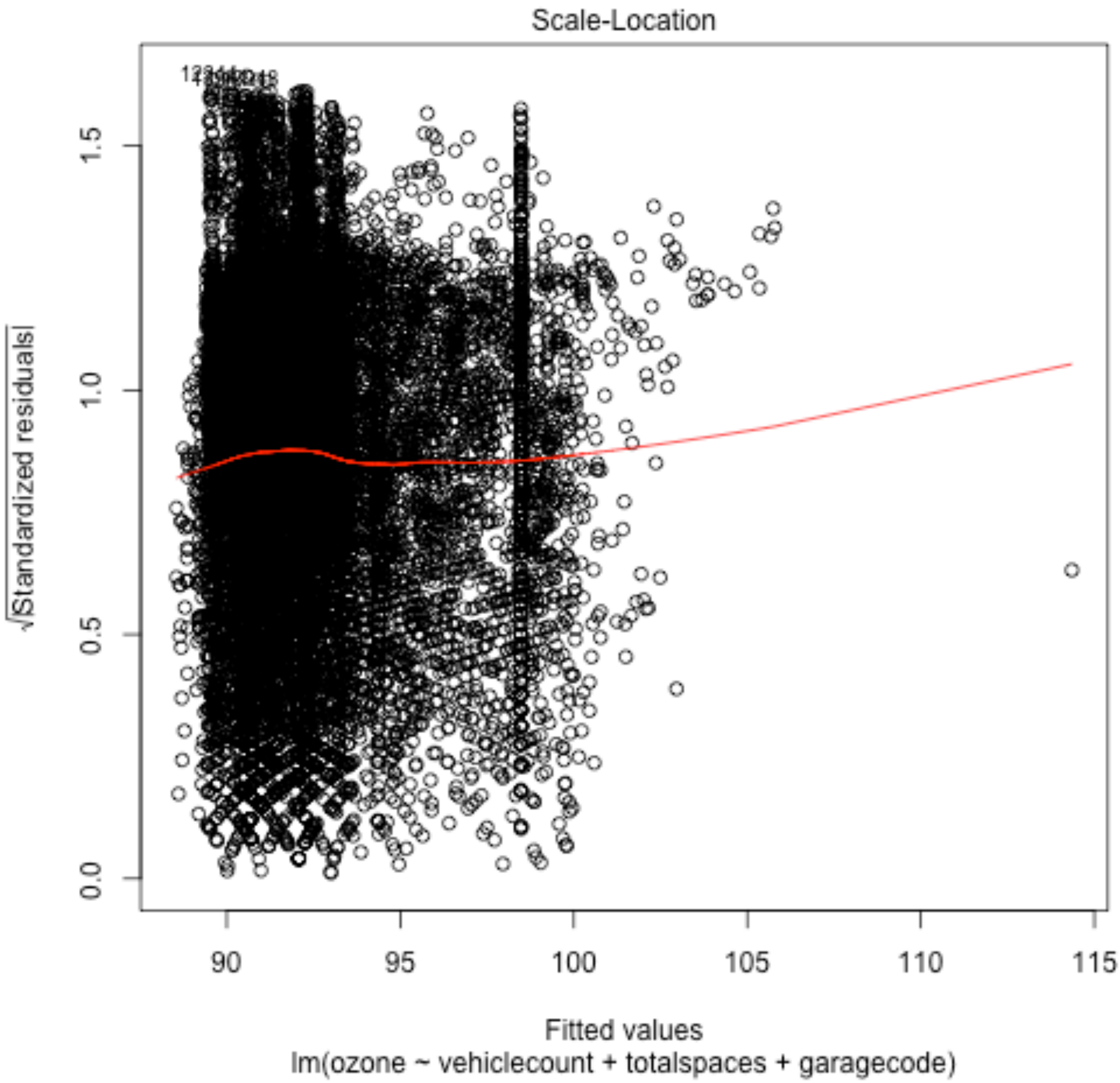
```
%r
plot(model)
```



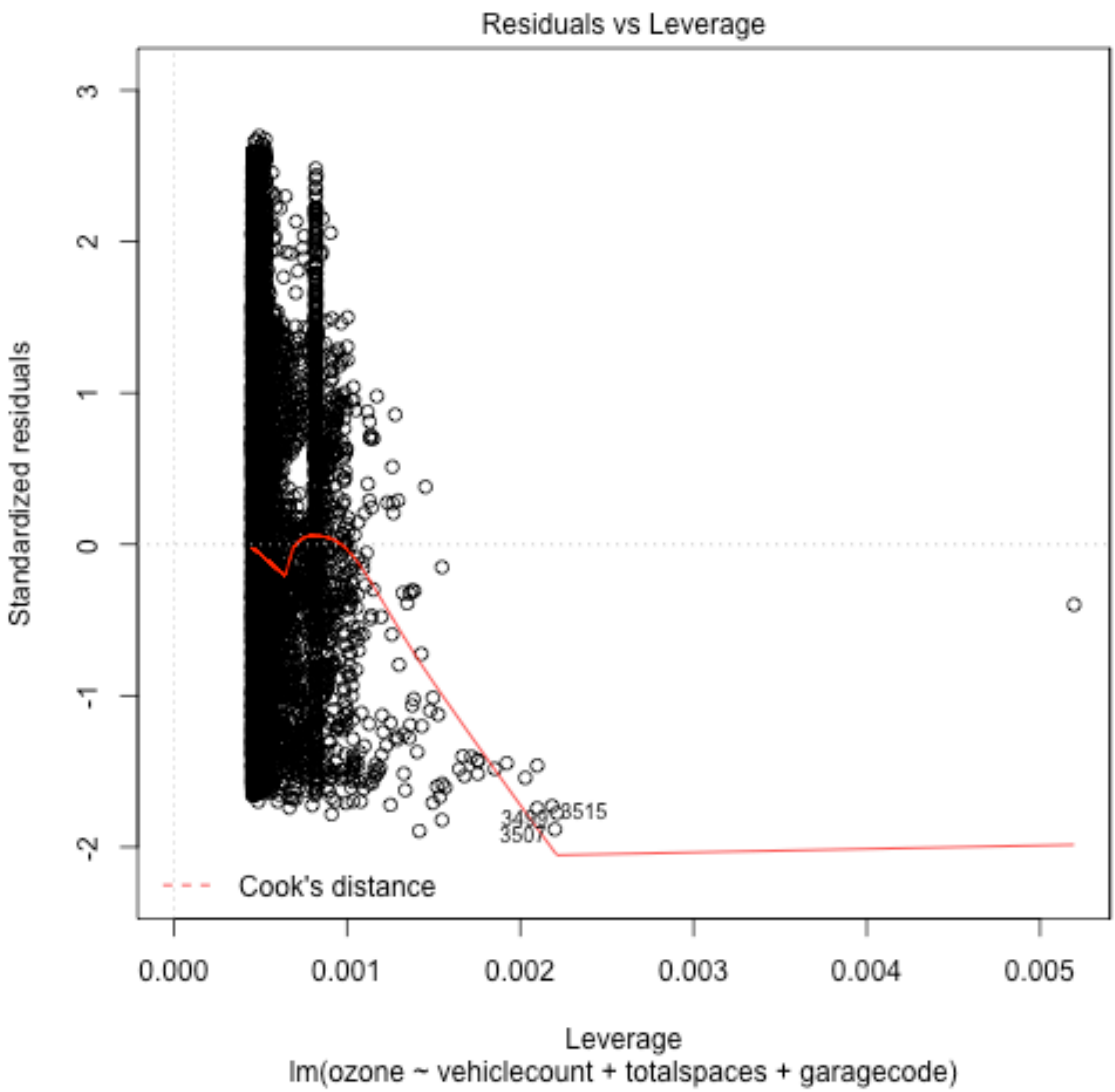
Residuals vs Fitted
Residuals
Fitted values
lm(ozone ~ vehiclecount + totalspaces + garagecode)
150
100
50
0
-50
-100
90
95
100
105
110
115



Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
-4
-2
0
2
4
Theoretical Quantiles
lm(ozone ~ vehiclecount + totalspaces + garagecode)



Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
90
95
100
105
110
115
Fitted values
lm(ozone ~ vehiclecount + totalspaces + garagecode)



Depicted the column names of the parking dataset

FINISHED

```
%r
colnames(aarhus_parking)
```

[1] "vehiclecount" "totalspaces" "garagecode" "ozone"

Depicted the structure of the parking dataset

FINISHED

```
%r
str(aarhus_parking)
```

```
'data.frame':   55264 obs. of  4 variables:
 $ vehiclecount: int  0 0 869 22 124 106 115 233 0 0 …
 $ totalspaces : int  65 512 1240 953 130 400 210 700 65 512 …
 $ garagecode  : Factor w/ 8 levels “BRUUNS”,“BUSGADEHUSET”,..: 5 8 7 1 2 4 3 6 5 8 …
 $ ozone       : int  101 106 107 103 105 106 110 106 106 110 …
```

Showed the summary of the parking datatset

FINISHED

```
%r
summary(aarhus_parking)
```

```
vehiclecount     totalspaces             garagecode        ozone<br />
 Min.   :   0.0   Min.   :  65.0   BRUUNS        : 6908   Min.   : 15.00<br />
 1st Qu.:  32.0   1st Qu.: 190.0   BUSGADEHUSET  : 6908   1st Qu.: 54.00<br />
 Median :  96.0   Median : 456.0   KALKVAERKSVEJ: 6908   Median : 87.00<br />
 Mean   : 192.2   Mean   : 526.2   MAGASIN       : 6908   Mean   : 92.42<br />
 3rd Qu.: 296.0   3rd Qu.: 763.2   NORREPORT     : 6908   3rd Qu.:127.00<br />
 Max.   :1464.0   Max.   :1240.0   SALLING       : 6908   Max.   :215.00<br />
                                   (Other)       :13816   NA's   :37696
```

Calling the lm function

FINISHED

```

%r
summary(lm(ozone~vehiclecount+totalspaces+garagecode, data = aarhus_parking))
```

```
Call:
lm(formula = ozone ~ vehiclecount + totalspaces + garagecode,
    data = aarhus_parking)
```

Showing the model

FINISHED

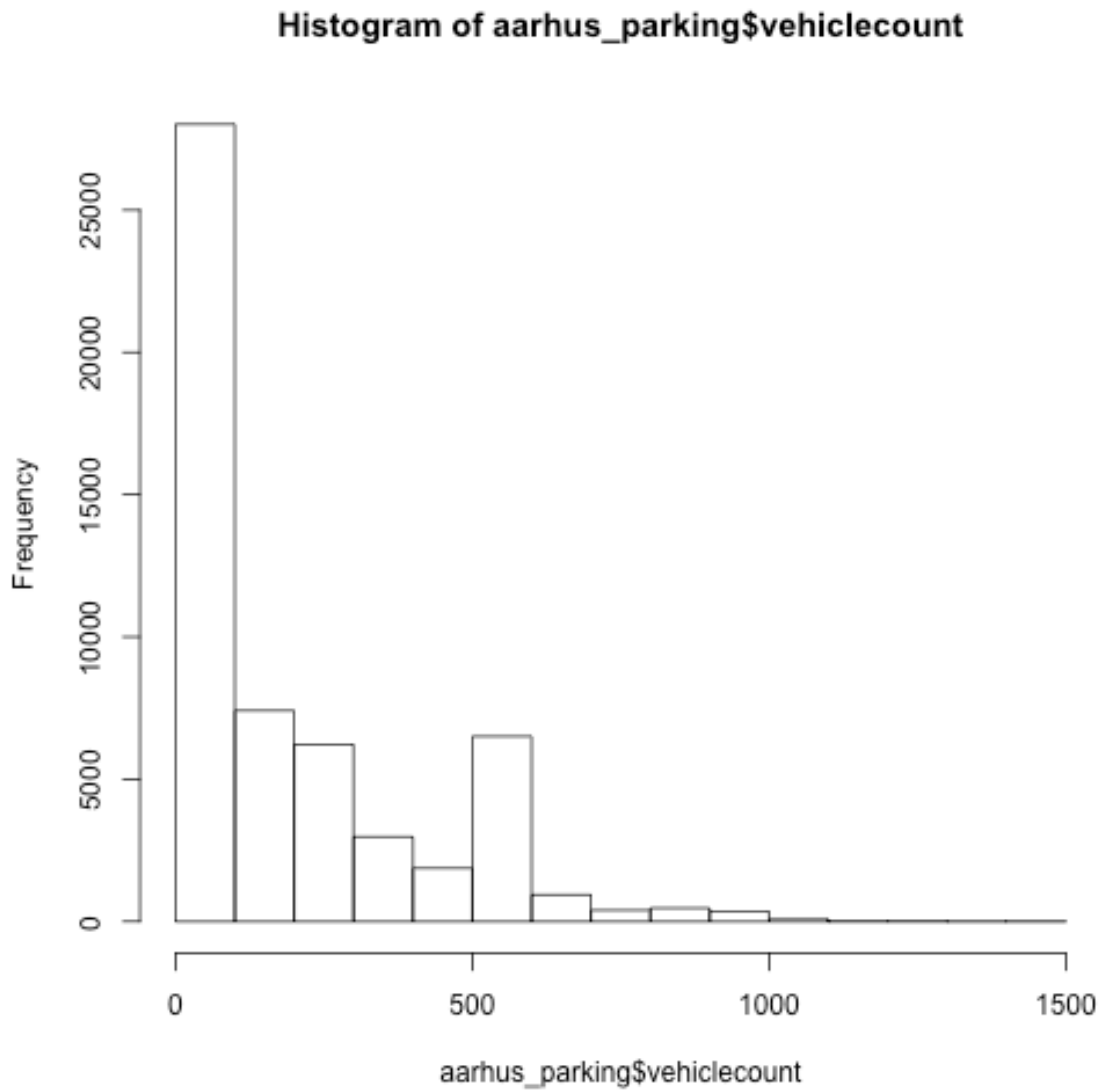
```
%r
model
```

```
Call:
lm(formula = ozone ~ vehiclecount + totalspaces + garagecode,
    data = aarhus_parking)
```

Histogram depicting the vehicle count

FINISHED

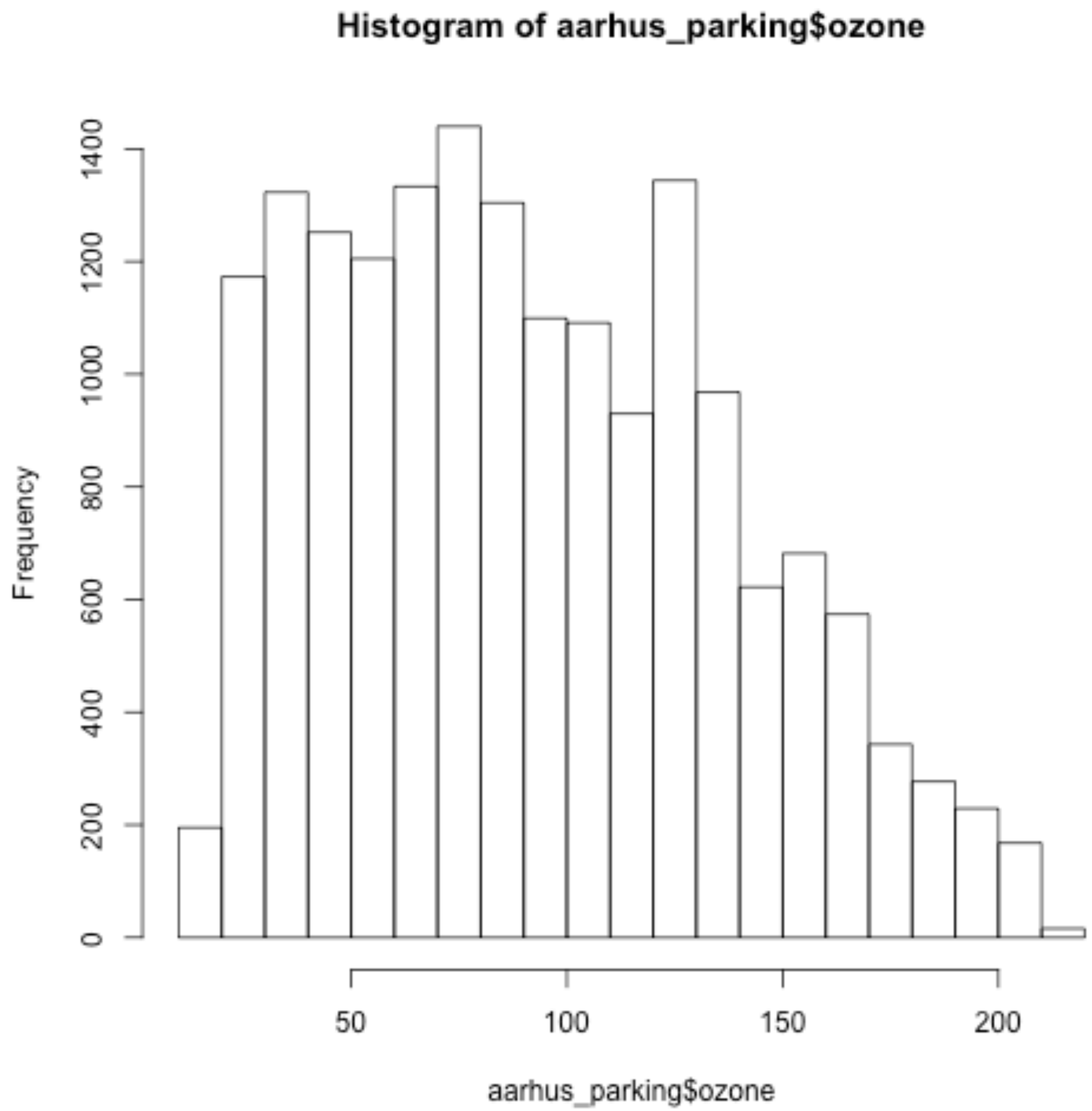
```
%r
hist(aarhus_parking$vehiclecount)
```



Histogram depicting the ozone layer

FINISHED

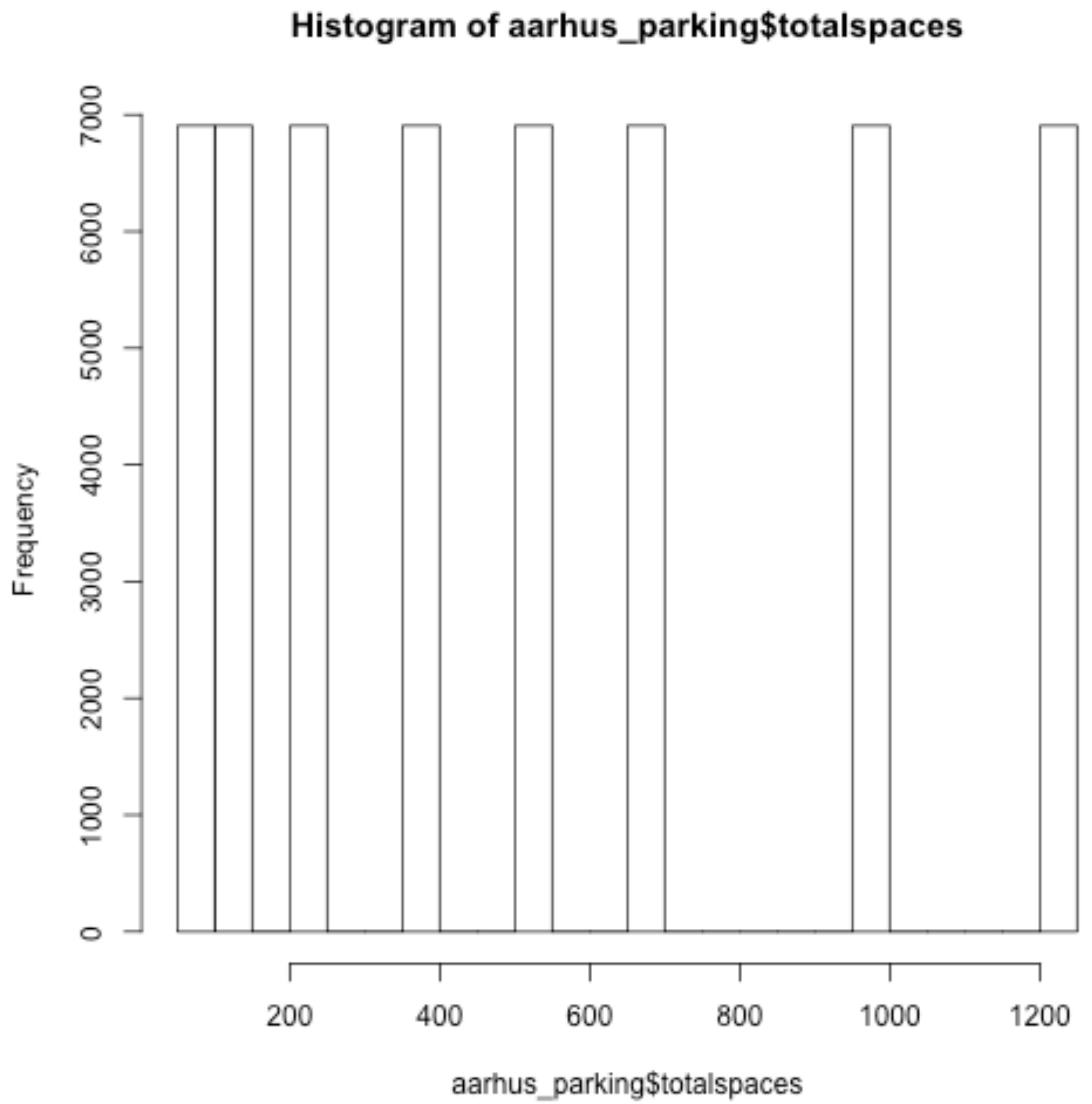
```
1 %r
2 hist(aarhus_parking$ozone)
```



Histogram depicting the total spaces

FINISHED

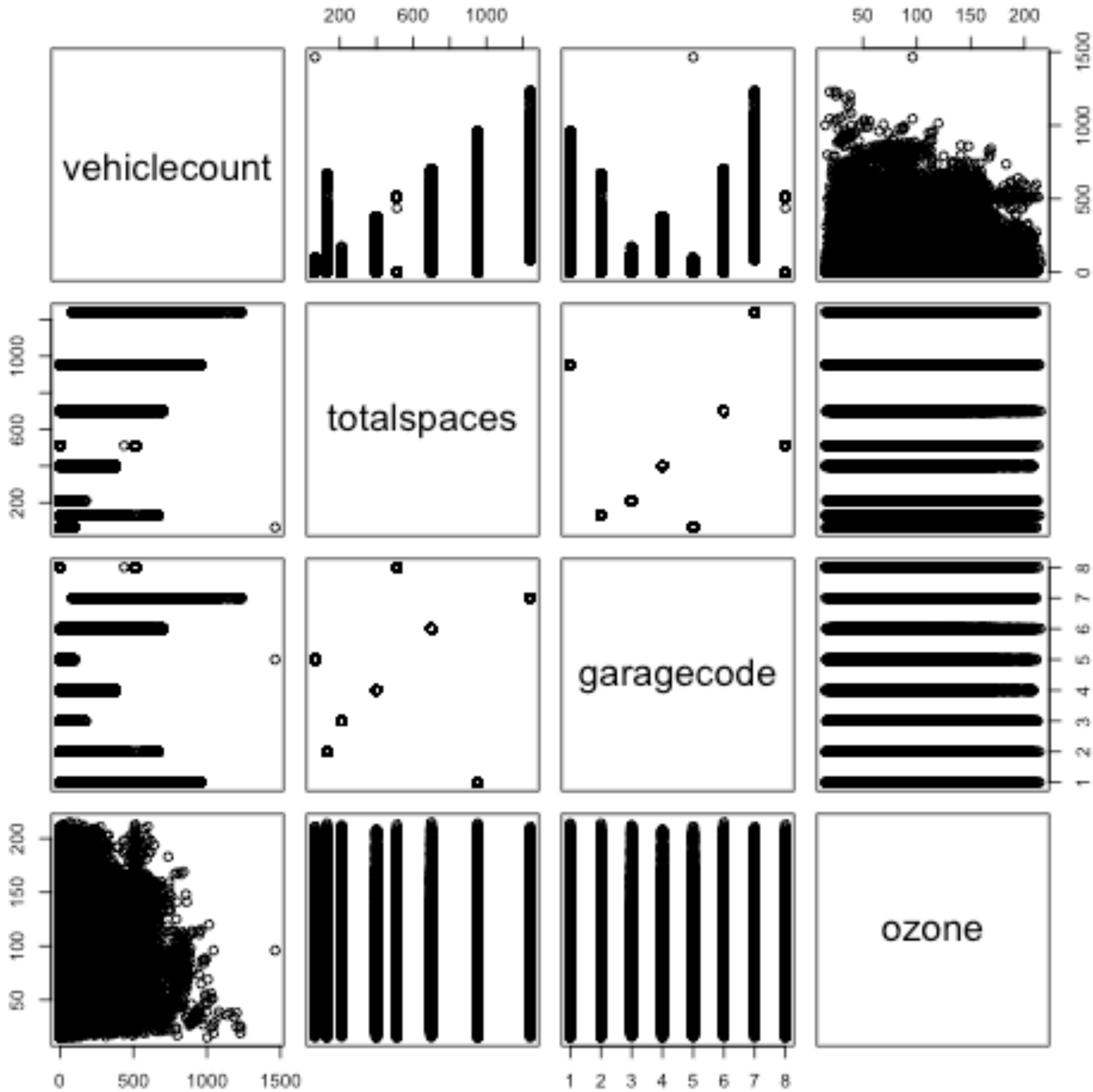
```
%r
hist(aarhus_parking$totalspaces)
```



Ggplot depicting the vehicle count, total spaces, garage code, and ozone layer

FINISHED

```
%r {"imageWidth":"400px}
library("ggplot2")
plot(aarhus_parking)
```



FINISHED

```
%spark.r
frequency(aarhus_parking)
```

```
[1] 1
```

Time series using the parking data set

FINISHED

```
%r
modelone <- ts(aarhus_parking, frequency=12, start=c(1946,1))
```

Printing the output for the the time series

FINISHED

```
%r
modelone
```

```
Jan 1946      0     65    5  101
Feb 1946      0    512    8  106
Mar 1946    869   1240    7  107
Apr 1946     22    953    1  103
May 1946    124    130    2  105
Jun 1946    106    400    4  106
Jul 1946    115    210    3  110
Aug 1946    233    700    6  106
Sep 1946      0     65    5  106
Oct 1946      0    512    8  110
Nov 1946    959   1240    7  115
Dec 1946     22    953    1  114
Jan 1947    124    130    2  118
Feb 1947    119    400    4  113
Mar 1947    121    210    3  114
Apr 1947    282    700    6  115
May 1947      0     65    5  115
Jun 1947      0    512    8  120
```

FINISHED

```
sc
```

```
res0: org.apache.spark.SparkContext = org.apache.spark.SparkContext@5add6c08
```

Created some partitions from the dataset

FINISHED

```
import org.apache.spark.mllib.util.LinearDataGenerator
val numRows = 10000
val numCols = 1000
val rawData = LinearDataGenerator.generateLinearRDD(sc, numRows, numCols, 1).toDF()
// Repartition into a more parallelism-friendly number of partitions
val data = rawData.repartition(64).cache()
```

```
import org.apache.spark.mllib.util.LinearDataGenerator
numRows: Int = 10000
numCols: Int = 1000
rawData: org.apache.spark.sql.DataFrame = [label: double, features: vector]
data: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [label: double, features: ve
ctor]
```

Prints out the coefficients from the model

FINISHED

```
%r
coefficients(model)
```

```
garagecodeBUSGADEHUSET garagecodeKALKVAERKSVEJ       garagecodeMAGASIN
          -2.191652364             0.238790550            -0.140165941
     garagecodeNORREPORT       garagecodeSALLING   garagecodeSCANDCENTER
           0.047225980            -0.504496446            -1.439290374
   garagecodeSKOLEBAKKEN
                    NA
```

Calculated the 95% confidence interval

FINISHED

```
%r
confint(model, level=0.95)
```

```
(Intercept)            87.491821077 96.959928651
vehiclecount            0.010913491  0.019515627
totalspaces            -0.009247282  0.003236974
garagecodeBUSGADEHUSET -6.649535257  2.266230530
garagecodeKALKVAERKSVEJ -3.766315284  4.243896385
garagecodeMAGASIN      -3.274337972  2.994006090
garagecodeNORREPORT    -4.711821775  4.806273736
garagecodeSALLING      -2.884713945  1.875721053
garagecodeSCANDCENTER  -5.393283802  2.514703053
garagecodeSKOLEBAKKEN            NA           NA
```

Fitted my model

FINISHED

```
%r
fitted(model)
```

```
92.07777  90.68724 100.28165  89.69668  91.53016  92.49639  93.58326
       8         9        10        11        12        13        14
 93.16276  92.07777  90.68724 101.65096  89.69668  91.53016  92.69418
       15        16        17        18        19        20        21
 93.67454  93.90828  92.07777  90.68724 102.48776  89.69668  91.54537
       22        23        24        25        26        27        28
 93.18105  93.73540  95.23194  89.69668  92.07777  90.68724 102.95941
       29        30        31        32        33        34        35
 91.83445  93.72877  93.79626  96.41868  92.07777  90.68724  93.34381
       36        37        38        39        40        41        42
 89.69668  92.15395  94.01785  92.79210  93.45184  92.07777  90.68724
       43        44        45        46        47        48        49
 92.53743  89.69668  91.98659  93.80484  92.77689  92.72154  92.07777
       50        51        52        53        54        55        56
 90.68724  92.11143  89.69668  92.18438  93.34841  92.74646  92.35639
       57        58        59        60        61        62        63
 92.07777  90.68724  91.94407  89.69668  92.16917  93.16583  92.67038
       64        65        66        67        68        69        70
```

Printed the residuals of the model

FINISHED

```
%r
residuals(model)
```

```
8.922234149  15.312763851   6.718354712  13.303316362  13.469842217
          6             7             8             9            10
 13.503609360  16.416742630  12.837237043  13.922234149  19.312763851
          11            12            13            14            15
 13.349044437  24.303316362  26.469842217  20.305820098  20.325455279
          16            17            18            19            20
 21.091723671  22.922234149  29.312763851  17.512243713  25.303316362
          21            22            23            24            25
 18.454627659  14.818954223  13.264597044   6.768057071  11.303316362
          26            27            28            29            30
 11.922234149  10.312763851  -6.959407604   4.165551045   6.271230113
          31            32            33            34            35
 10.203738810   5.581321499   6.922234149  10.312763851  10.656193441
          36            37            38            39            40
 18.303316362  11.846045314   9.982153499  15.207899678  14.548160429
          41            42            43            44            45
 11.922234149   9.312763851  11.462565048  15.303316362  13.013405459
```

Getting the analysis of the variance table

FINISHED

```
1 %r
2 anova(model)
```

```
Analysis of Variance Table
```

Calculated the variance-covariance of the model

FINISHED

```
1 %r
2 vcov(model)
```

```
(Intercept)               5.833246e+00 -6.449969e-05 -7.288449e-03
vehiclecount             -6.449969e-05  4.815003e-06 -9.173920e-07
totalspaces              -7.288449e-03 -9.173920e-07  1.014167e-05
garagecodeBUSGADEHUSET   -4.873959e+00 -6.963061e-04  6.137709e-03
garagecodeKALKVAERKSVEJ  -4.299779e+00  4.120483e-05  5.199843e-03
garagecodeMAGASIN        -2.911121e+00 -7.214983e-05  3.327733e-03
garagecodeNORREPORT      -5.358439e+00  4.512334e-05  6.644293e-03
garagecodeSALLING        -7.188819e-01 -2.227684e-04  3.663671e-04
garagecodeSCANDCENTER     3.227003e+00 -4.830274e-04 -4.966160e-03
                         garagecodeBUSGADEHUSET garagecodeKALKVAERKSVEJ
(Intercept)                       -4.8739590503           -4.299779e+00
vehiclecount                      -0.0006963061            4.120483e-05
totalspaces                        0.0061377092            5.199843e-03
garagecodeBUSGADEHUSET             5.1725089901            3.616269e+00
garagecodeKALKVAERKSVEJ            3.6162686153            4.175149e+00
garagecodeMAGASIN                  2.4917027598            2.215533e+00
garagecodeNORREPORT                4.4864332678            3.961114e+00
```

FINISHED

Checks for the quality of the regression fits

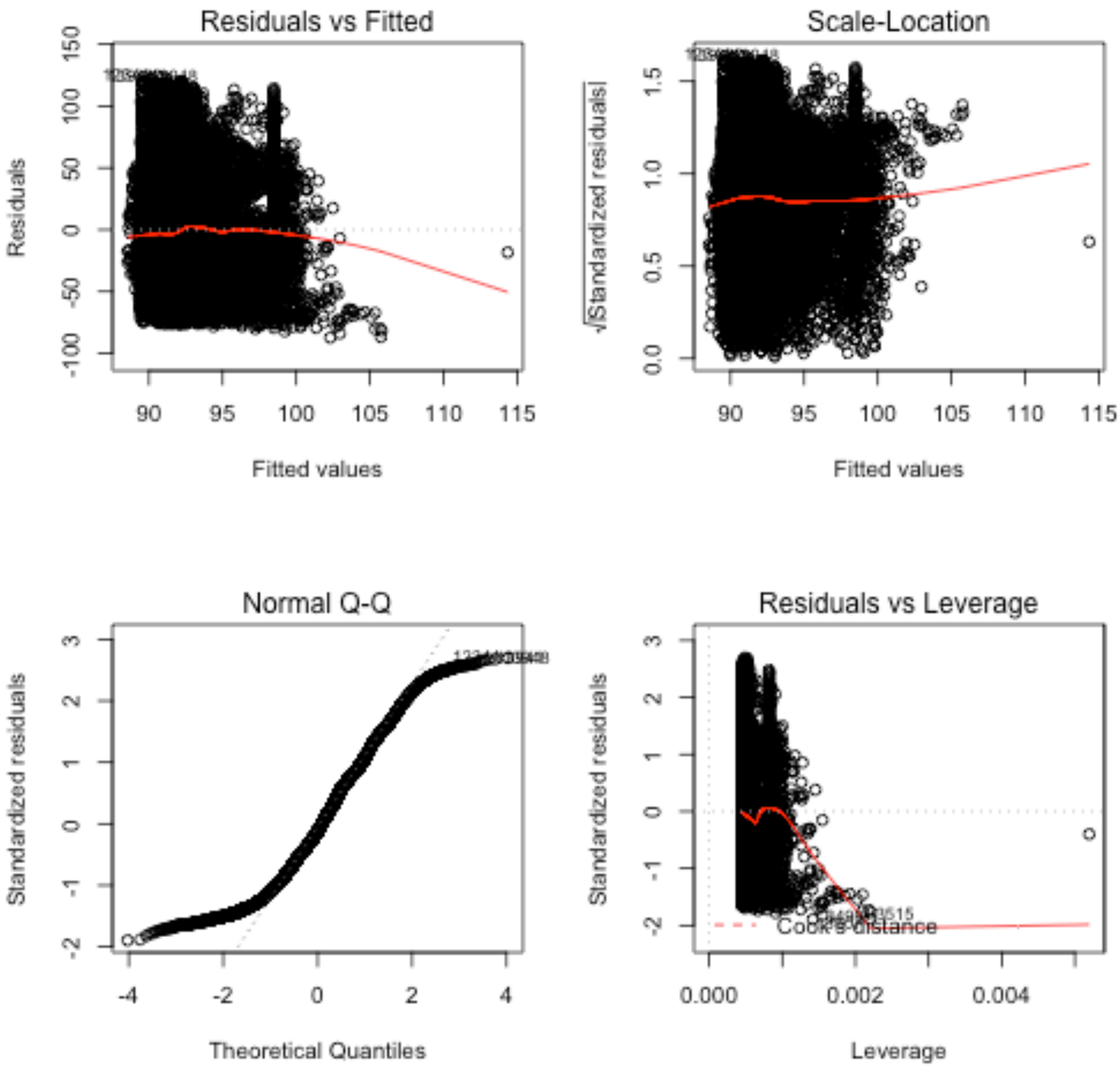
```
%r
influence(model)
```

```
$hat
           1            2            3            4            5 
0.0004559825 0.0004832205 0.0010648366 0.0005230584 0.0004631889 
           6            7            8            9           10 
0.0004553779 0.0004665068 0.0004589876 0.0004559825 0.0004832205 
          11           12           13           14           15 
0.0012945228 0.0005230584 0.0004631889 0.0004558431 0.0004684927 
          16           17           18           19           20 
0.0004732811 0.0004559825 0.0004832205 0.0014529287 0.0005230584 
          21           22           23           24           25 
0.0004629252 0.0004602460 0.0004699072 0.0005254274 0.0005230584 
          26           27           28           29           30 
0.0004559825 0.0004832205 0.0015482434 0.0004587750 0.0004707375 
          31           32           33           34           35 
0.0004713940 0.0006012955 0.0004559825 0.0004832205 0.0004643620 
          36           37           38           39           40 
0.0005230584 0.0004560883 0.0004786388 0.0004561188 0.0004632405 
          41           42           43           44           45 
```

FINISHED

Shows the different plots

```
%r
layout(matrix(c(1,2,3,4),2,2))
plot(model)
```



Installed the Data Analysis and Graphics package

FINISHED

```
1 %r
2 install.packages("DAAG", repos = "http://cran.us.r-project.org")
```

```
The downloaded binary packages are in
    /var/folders/ll/1mpcgfrd7nlgpz03y3z75t6w0000gn/T//RtmptTl8YT/downloaded_packages
```

Installed the bootstrap package

FINISHED

```
1 %r
```

```
The downloaded binary packages are in
    /var/folders/ll/1mpcgfrd7nlgpz03y3z75t6w0000gn/T//RtmptTl8YT/downloaded_packages
```

Defined the functions

FINISHED

```
%r
library(bootstrap)
theta.model <- function(x,y){lsmodel(x,y)}
theta.predict <- function(model,x){cbind(1,x)%*%model$coef}
```

Converted the data frame to a numeric matrix

FINISHED

```
%r
X <- as.matrix(model[c("ozone","vehiclecount","totalspaces")])
y <- as.matrix(model[c("garagecode")])
```

Installed the MASS package

FINISHED

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

```
The downloaded binary packages are in
    /var/folders/ll/1mpcgfrd7nlgpz03y3z75t6w0000gn/T//RtmptTl8YT/downloaded_packages
```

Performed a stepwise model selection by AIC

FINISHED

```
%r
library(MASS)
modelfit <- lm(ozone~vehiclecount+totalspaces+garagecode,data=aarhus_parking)
step <- stepAIC(model, direction="both")
step$anova
```

```
Start:  AIC=134634.2
ozone ~ vehiclecount + totalspaces + garagecode
```

Installed the leaps package

FINISHED

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

```
There is a binary version available (and will be installed) but
  the source version is later:
      binary source
leaps    2.9    3.0
```

FINISHED

Used the "leaps" function to get the best subsets of the variables

```
%r
library(leaps)
attach(aarhus_parking)
leaps<-regsubsets(ozone~vehiclecount+totalspaces+garagecode,data=aarhus_parking,nbest=
```

Printed the subset selection

FINISHED

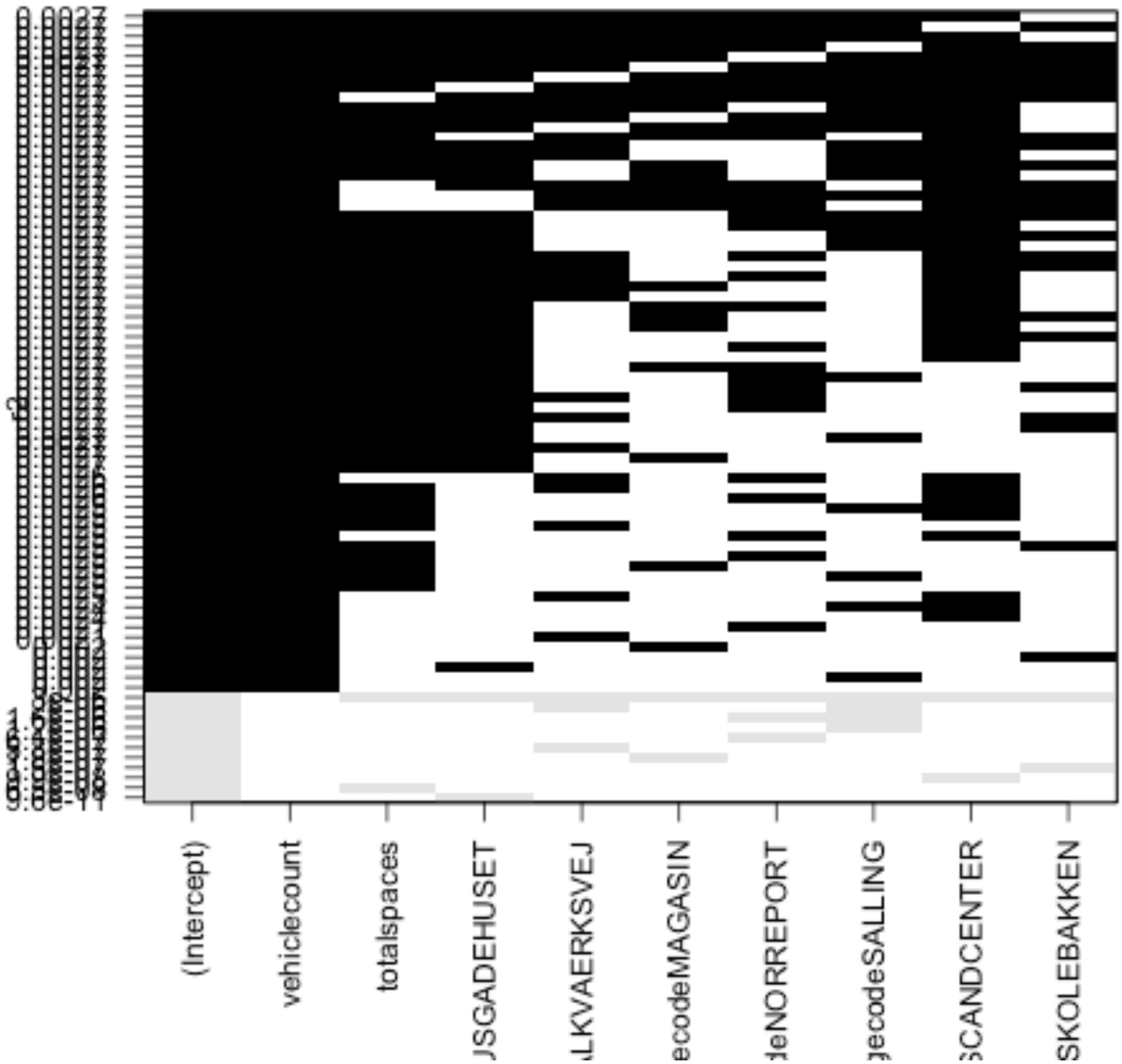
```
%r
summary(leaps)
```

```
Subset selection object
Call: regsubsets.formula(ozone ~ vehiclecount + totalspaces + garagecode,
    data = aarhus_parking, nbest = 10)
9 Variables  (and intercept)
                        Forced in Forced out
vehiclecount                FALSE      FALSE
totalspaces                 FALSE      FALSE
garagecodeBUSGADEHUSET      FALSE      FALSE
garagecodeKALKVAERKSVEJ     FALSE      FALSE
garagecodeMAGASIN           FALSE      FALSE
garagecodeNORREPORT         FALSE      FALSE
garagecodeSALLING           FALSE      FALSE
garagecodeSCANDCENTER       FALSE      FALSE
garagecodeSKOLEBAKKEN       FALSE      FALSE
10 subsets of each size up to 8
Selection Algorithm: exhaustive
          vehiclecount totalspaces garagecodeBUSGADEHUSET
1  ( 1 )  "<em>"          " "         " "<br />
```

Plotted the leaps model

FINISHED

```
%r
plot(leaps,scale="r2")
```



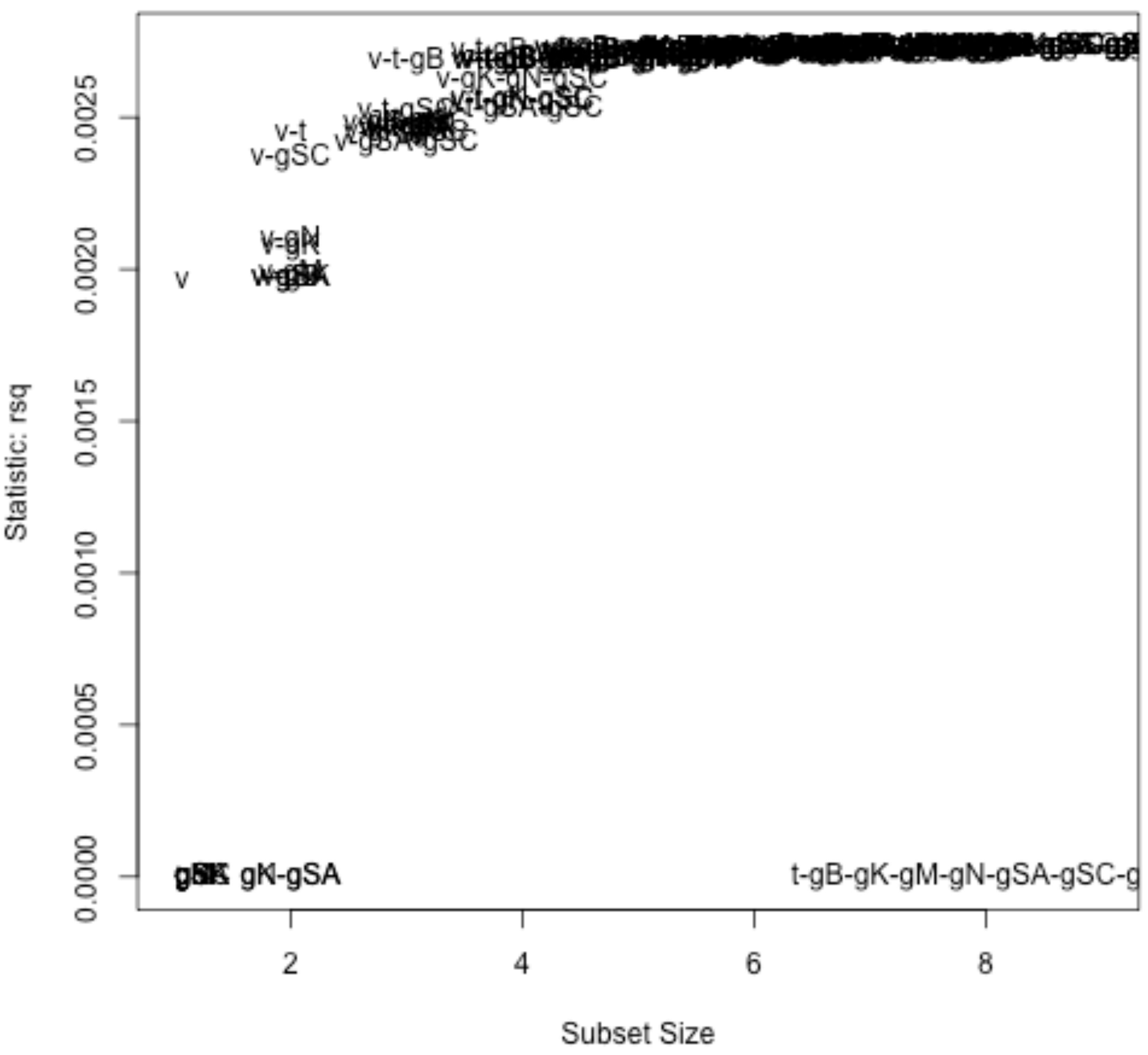
Installed the car package and got the subsets

FINISHED

```
%r
install.packages("car", repos = "http://cran.us.r-project.org")
library(car)
subsets(leaps, statistic="rsq")
```

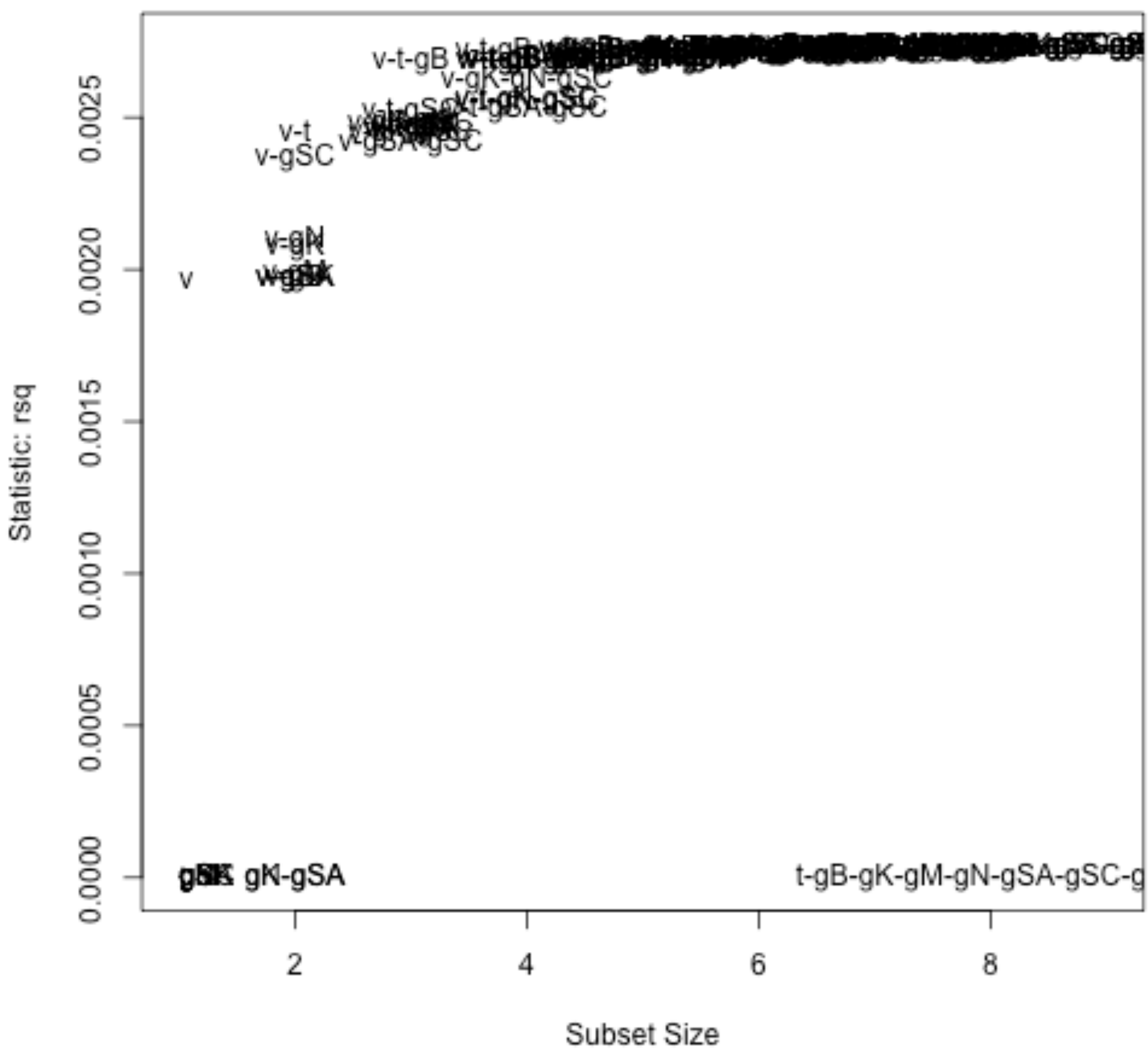
The downloaded binary packages are in
/var/folders/ll/1mpcgfrd7nlgpz03y3z75t6w0000gn/T//RtmptTl8YT/downloaded_packages

```
Error in legend(if (!is.na(charmatch(legend[1], "interactive"))) locator(1) els
e if (is.character(legend)) legend else if (is.numeric(legend) && : invalid coo
rdinate lengths
```



FINISHED

```
Error in legend(if (!is.na(charmatch(legend[1], "interactive"))) locator(1) els
e if (is.character(legend)) legend else if (is.numeric(legend) && : invalid coo
rdinate lengths
```



Installed the caret package

FINISHED

```
%r
library(caret)
```

Defined the parking dataset

FINISHED

```
%r
data(aarhus_parking)
```

FINISHED

```
%r
train_control <- trainControl(method="cv", number=10)
```

FINISHED

```
%r
# fix the parameters of the algorithm
grid <- expand.grid(.fL=c(0), .usekernel=c(FALSE))
```

FINISHED

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

```
The downloaded binary packages are in
    /var/folders/ll/1mpcgfrd7nlgpz03y3z75t6w0000gn/T//RtmpgvXzmi/downloaded_packages
```

FINISHED

```
%spark.r
aarhus_parking <- read.csv("/Users/joannariascos/Desktop/algorithm/aarhus_parking.csv'
```

FINISHED

```
%r
colnames(aarhus_parking)
```

```
[1] “vehiclecount” “totalspaces”  “garagecode”   “ozone”
```

FINISHED

```
%r
na.omit(aarhus_parking)
```

1	0	65	NORREPORT	101
2	0	512	SKOLEBAKKEN	106
3	869	1240	SCANDCENTER	107
4	22	953	BRUUNS	103
5	124	130	BUSGADEHUSET	105
6	106	400	MAGASIN	106
7	115	210	KALKVAERKSVEJ	110
8	233	700	SALLING	106
9	0	65	NORREPORT	106
10	0	512	SKOLEBAKKEN	110
11	959	1240	SCANDCENTER	115
12	22	953	BRUUNS	114
13	124	130	BUSGADEHUSET	118
14	119	400	MAGASIN	113
15	121	210	KALKVAERKSVEJ	114
16	282	700	SALLING	115
17	0	65	NORREPORT	115
18	0	512	SKOLEBAKKEN	120

FINISHED

```
%r
model <- train(ozone~vehiclecount, data=aarhus_parking, trControl=train_control, method="nl
```

```
Error in train.default(x, y, weights = w, ...): wrong model type for regression
```

READY