Co2 Car Emissions

Javier Merino

2024-11-22

Contents

1 Objective

Build a linear regression model to predict a car's CO2 Emissions (grams / Km). The model will be built by using the "Fuel consumption ratings" dataset from the Government of Canada.

2 Project Overview

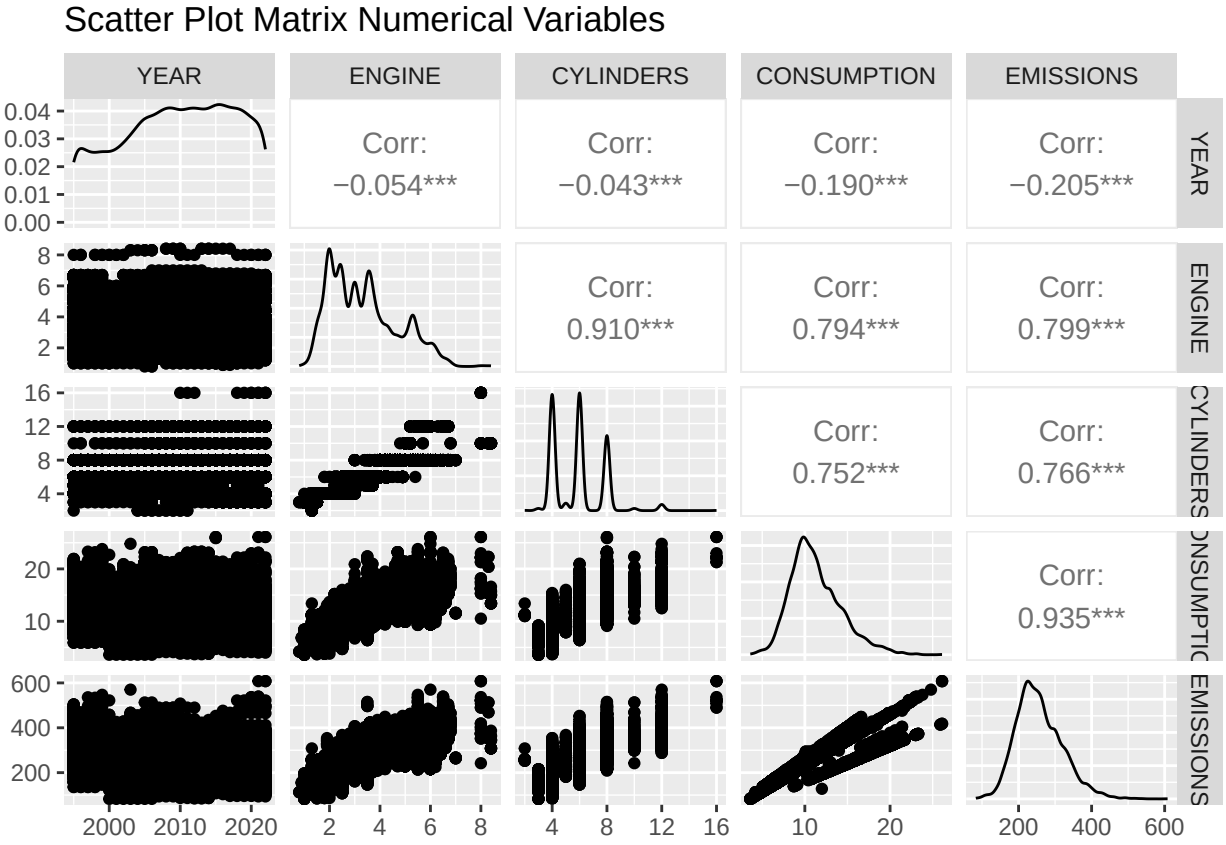
1. Made a descriptive analysis to understand the data and decide which variables could be dropped to proceed with the modelling section.
2. Split-ed the data into 80% for training and 20% for testing.
3. Applied the variable screening methods to the train data, and all methods suggested a full model which though high in r-squared value it had a pattern in the residual plot. The pattern was eliminated with an interaction and named the model as "Model 8".
4. Ran some models based on the descriptive analysis, and chose to use a simple model which still had high accuracy, and named the model as "Model 9".
5. Fitted the test data into models 8 and 9 and compared the prediction for the two models.

3 Descriptive Analysis

3.1 Numerical variables

- YEAR
- ENGINE
- CYLINDERS
- CONSUMPTION
- EMISSIONS (Response Variable)

3.1.1 Correlation Matrix



Insights

EMISSIONS has three significant correlations with the other three numerical variables:

- Emissions vs Fuel Consumption: 0.935
- Emissions vs Engine: 0.799
- Emissions vs Cylinders: 0.766

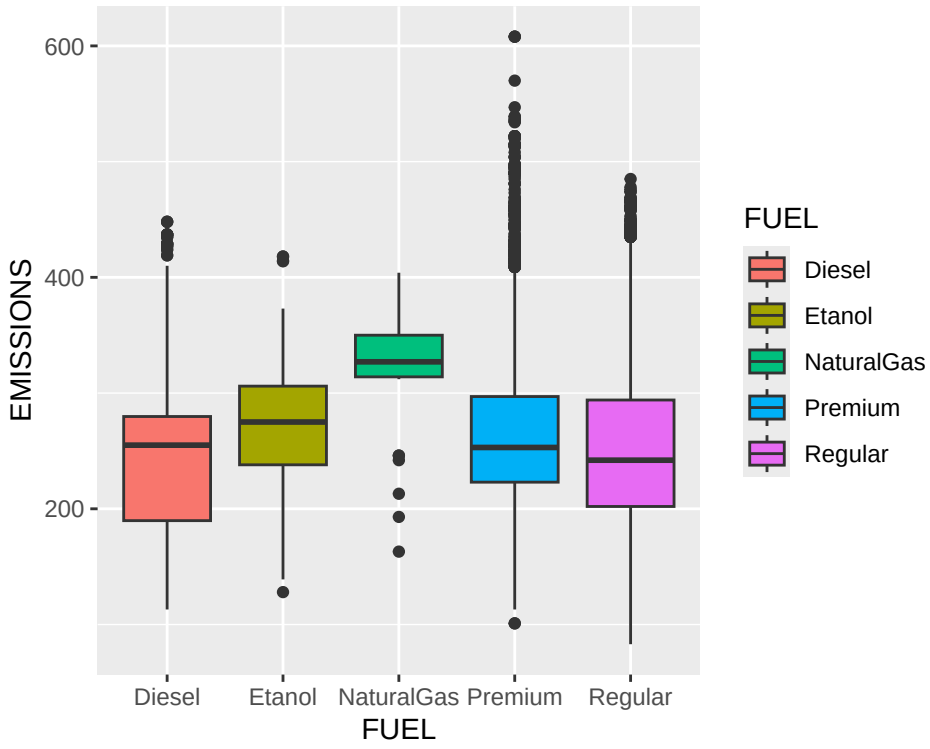
CYLINDERS has a high correlation with ENGINE: 0.910

3.1.2 Categorical Variables

- Brand: 55 levels.
- Model: 4185 levels.
- Class: 17 levels.
- Transmission: 5 levels.
- Fuel: 5 levels.
- Gears: 9 levels.

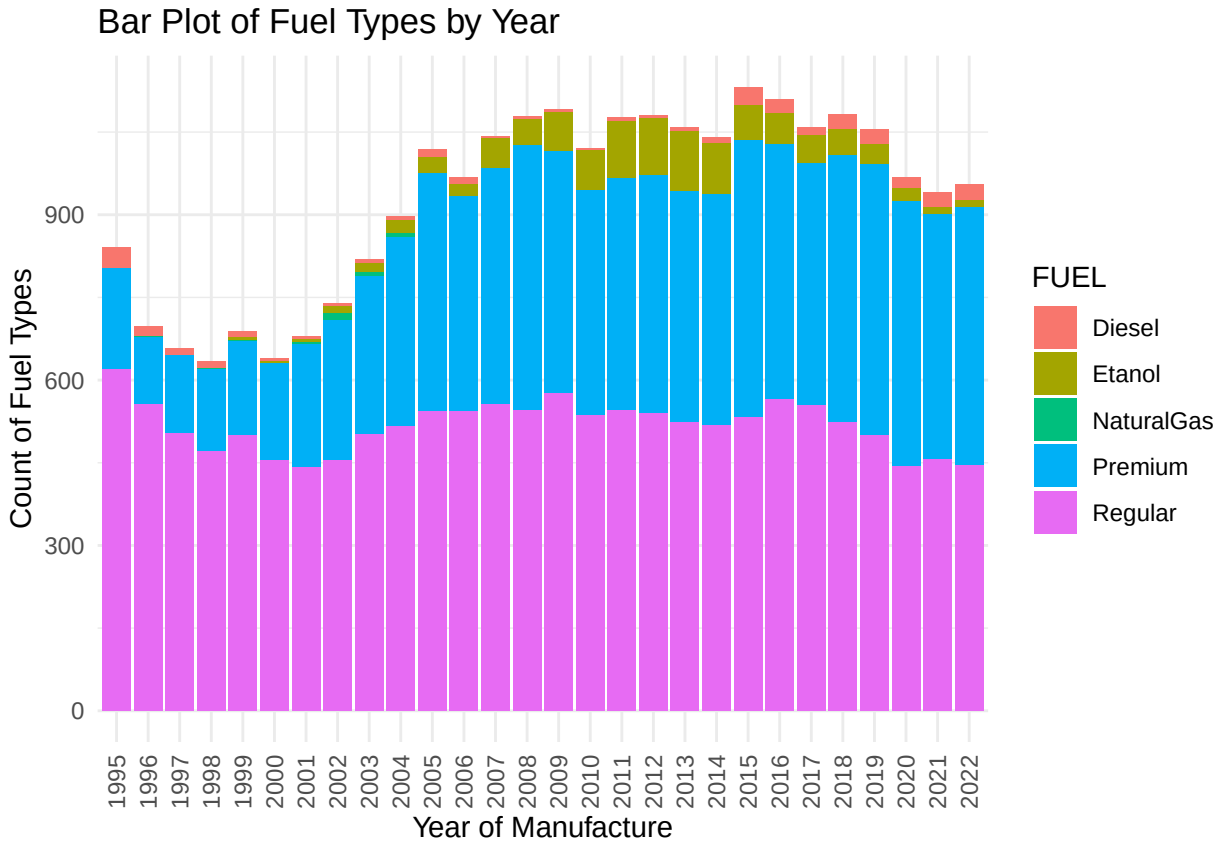
Did not considered "Car Model" in the Analysis, as this variables has a large number of categories, and it not provide useful information to predict emissions.

3.1.2.1 FUEL



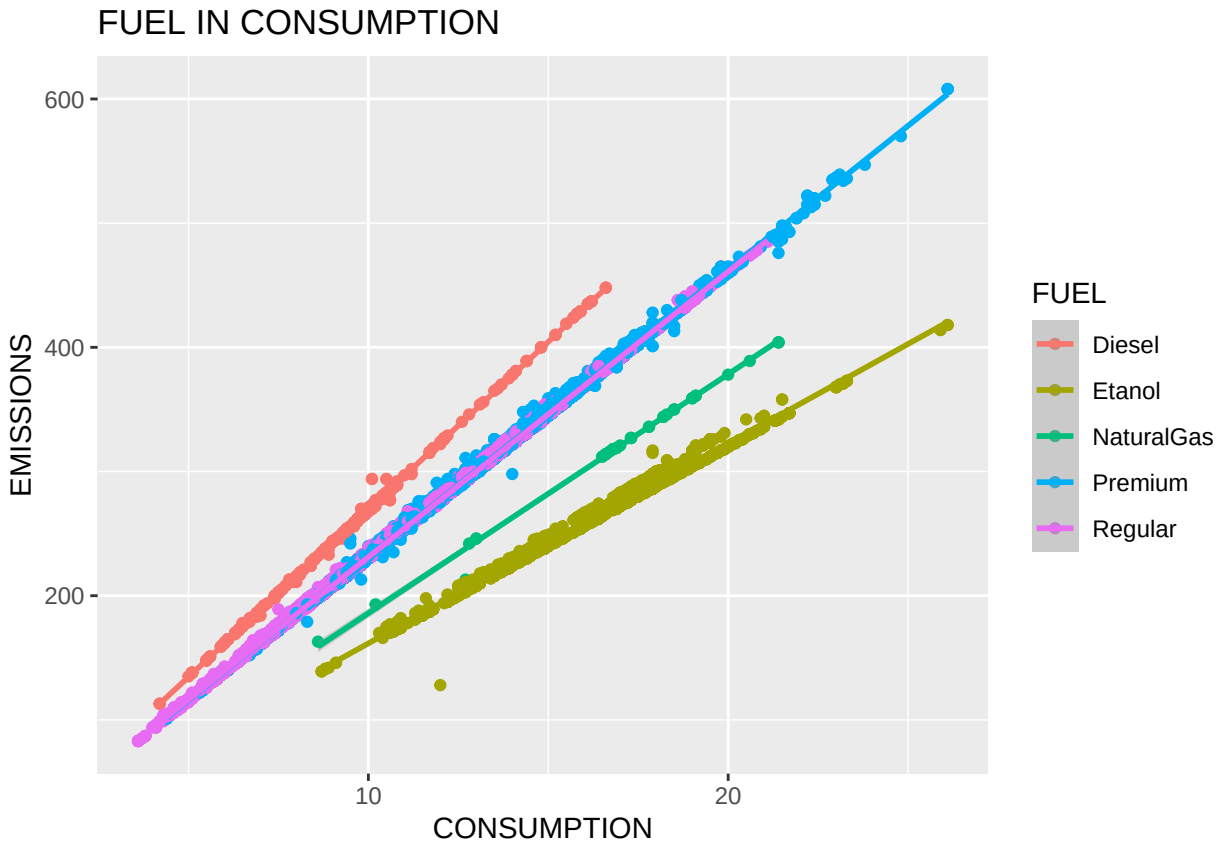
Insights:

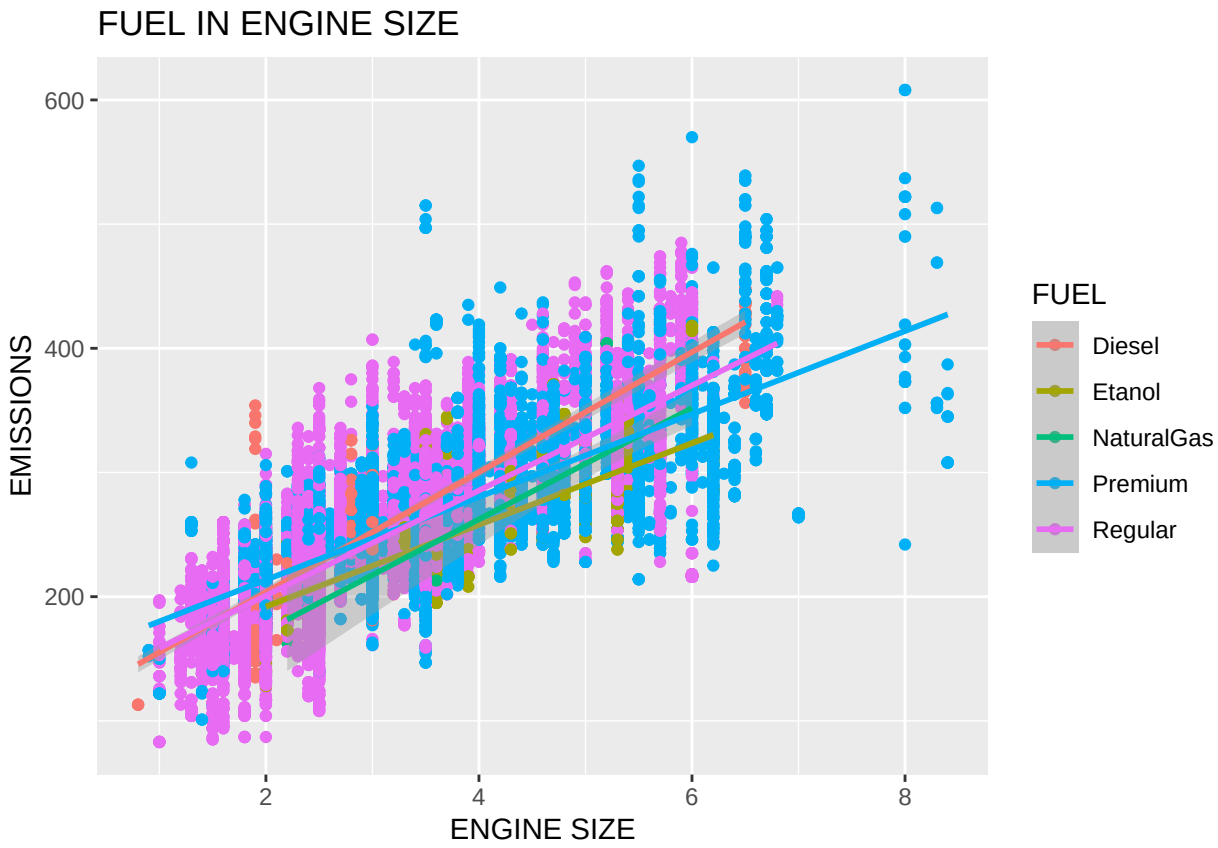
- The median CO2 emission for premium, regular and diesel are very close from each other, and they are surpassed by Ethanol and Natural Gas.

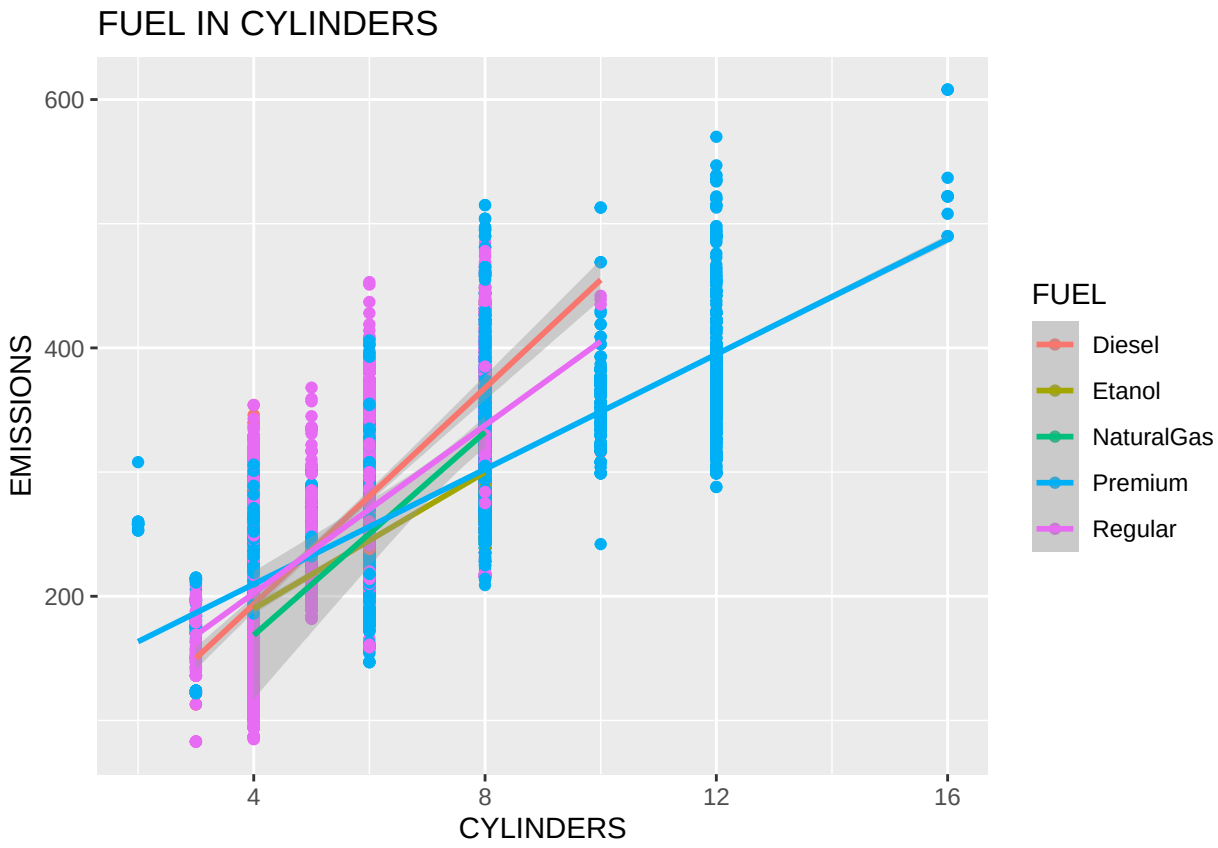


Insights:

- The majority of cars in the data set uses gasoline, which can be premium or regular in similar proportions. Ethanol grew in 2007 but the number was reduced by 2015.



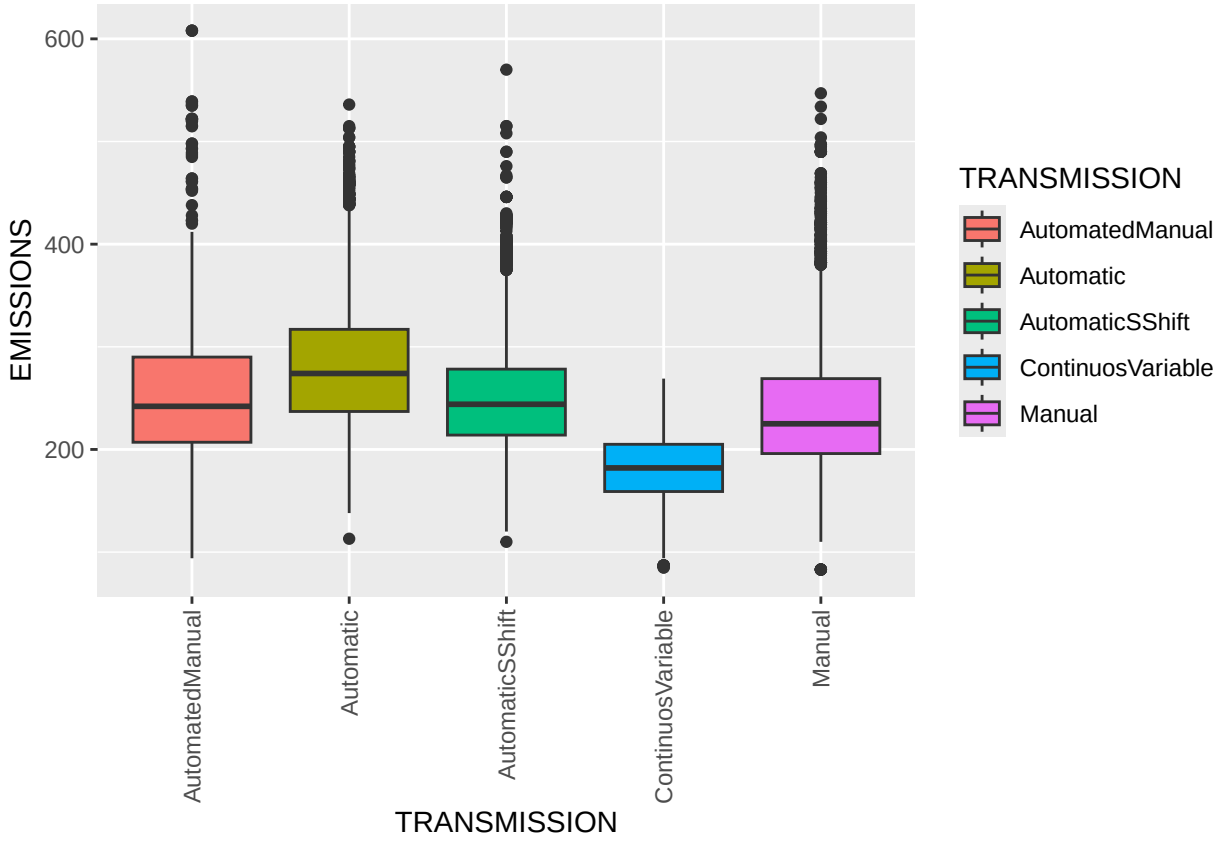

FUEL IN ENGINE SIZE
EMISSIONS
600
400
200
2
4
6
8
ENGINE SIZE
FUEL
Diesel
Etanol
NaturalGas
Premium
Regular


FUEL IN CYLINDERS
EMISSIONS
600
400
200
4
8
12
16
CYLINDERS
FUEL
Diesel
Etanol
NaturalGas
Premium
Regular

Insights:

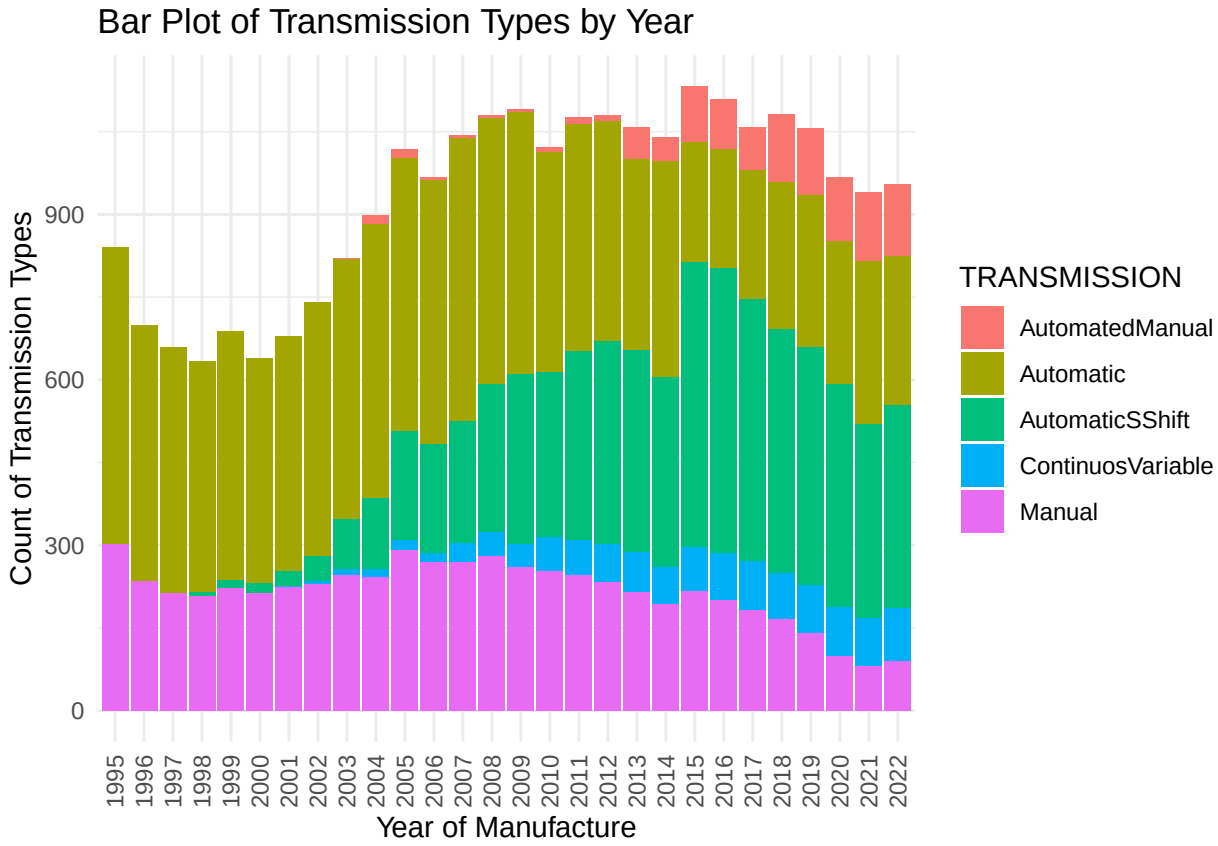
- The majority of cars uses gasoline and in consumption they are concentrated in the same line, so they could be considered as one type of gasoline. The lines are separated from each other, which makes them the perfect candidate for interaction, CONSUMPTION*FUEL
- The bigger the engine size, the gasoline premium is more required.
- From 10 cylinders and above the only fuel used is premium.

3.1.2.2 TRANSMISSION



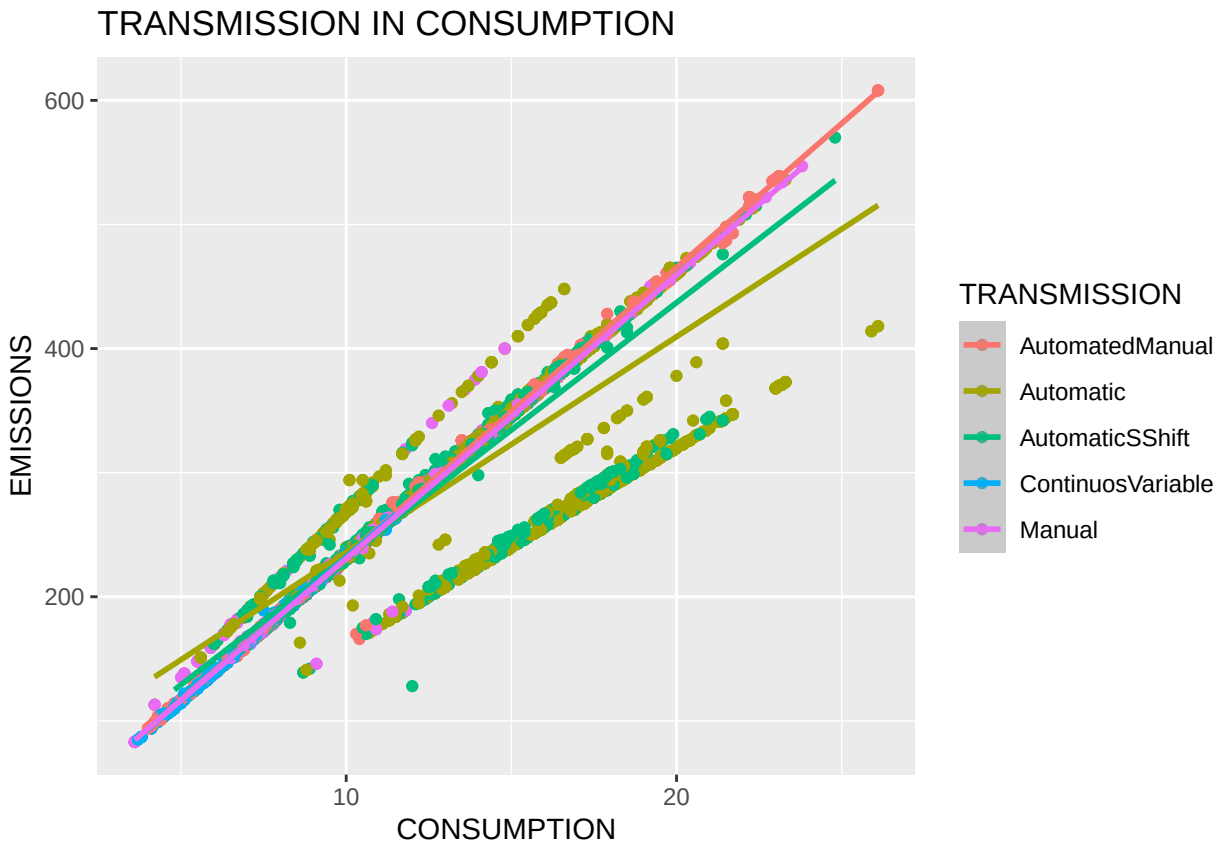
Insights:

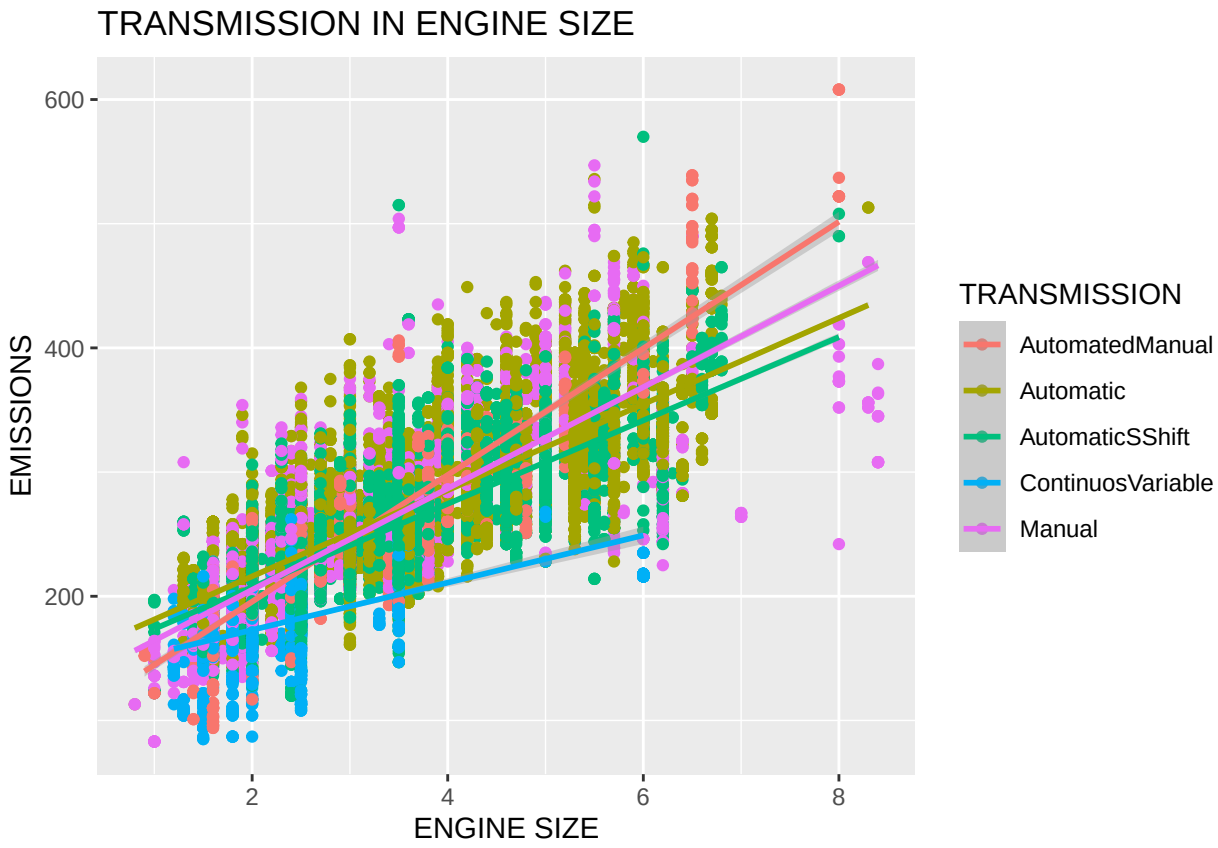
- Continuous variable has the lowest co2 emission. This type of automatic transmission uses pulleys and a steel band instead of traditional fixed gears. The CVT can change the gear ratio forever to maintain the engine running at peak efficiency. On the whole, the more gears offered in a typical automatic transmission, the better engine power is optimized.

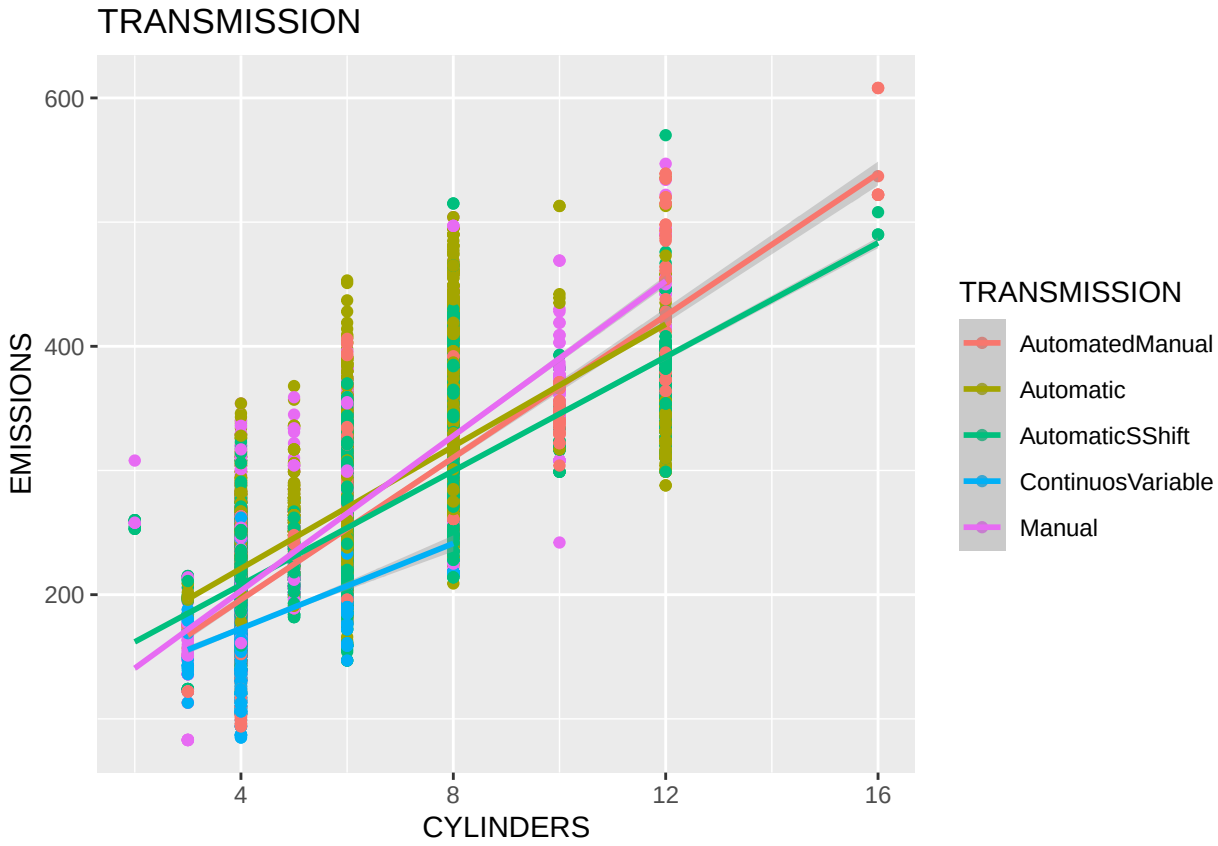


Insights:

- Automatic with selected Shift transmission grew considerably since 2005, and now is the main offer in the market. This kind of transmission allows to choose between fully automatic shifting or semi-automatic, clutch-less shifting.


TRANSMISSION IN CONSUMPTION
600
400
200
EMISSIONS
10
20
CONSUMPTION
TRANSMISSION
AutomatedManual
Automatic
AutomaticSShift
ContinuosVariable
Manual


TRANSMISSION IN ENGINE SIZE
600
400
200
EMISSIONS
2
4
6
8
ENGINE SIZE
TRANSMISSION
AutomatedManual
Automatic
AutomaticSShift
ContinuosVariable
Manual

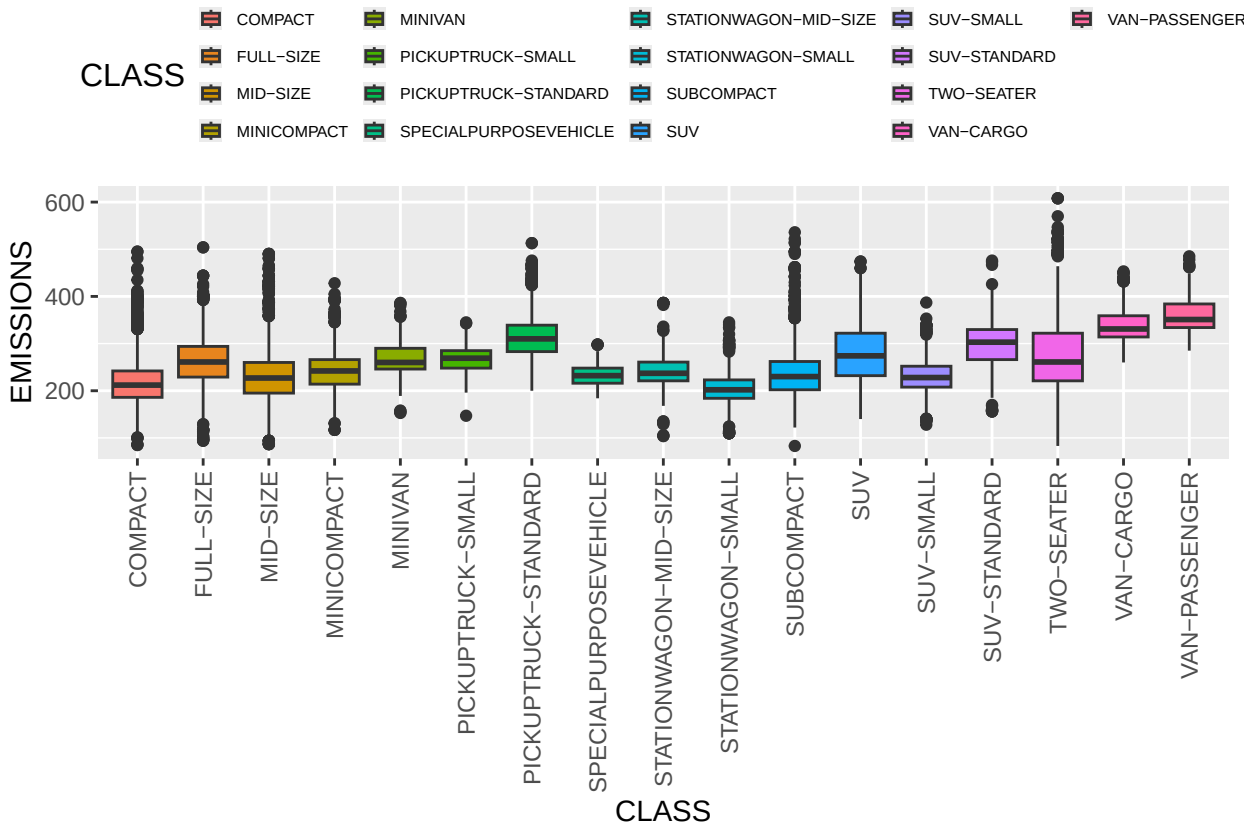


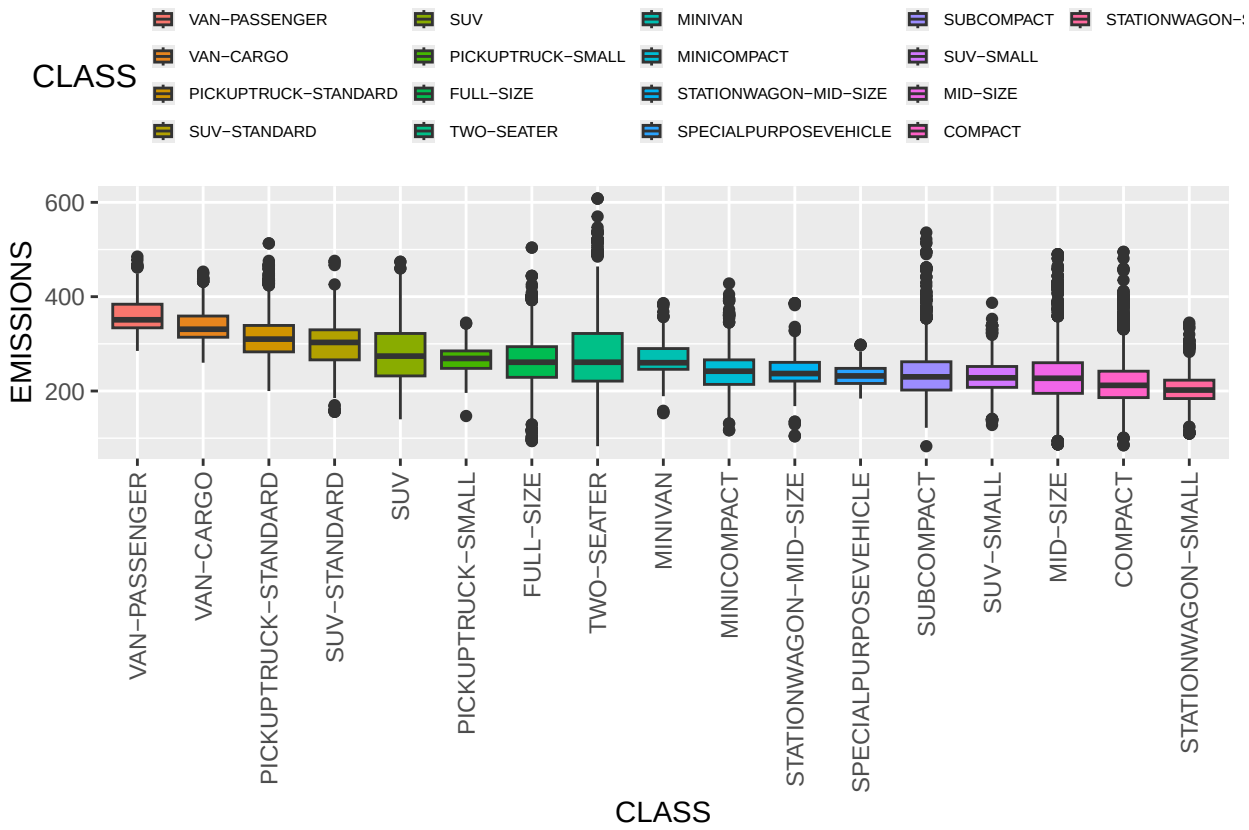
Insights:

- In consumption vs emissions, all the transmission lines are close to each other, which means that all transmission types are dispersed.
- In engine size vs emissions, all the transmission lines are close to each other.

3.1.2.3 CLASS

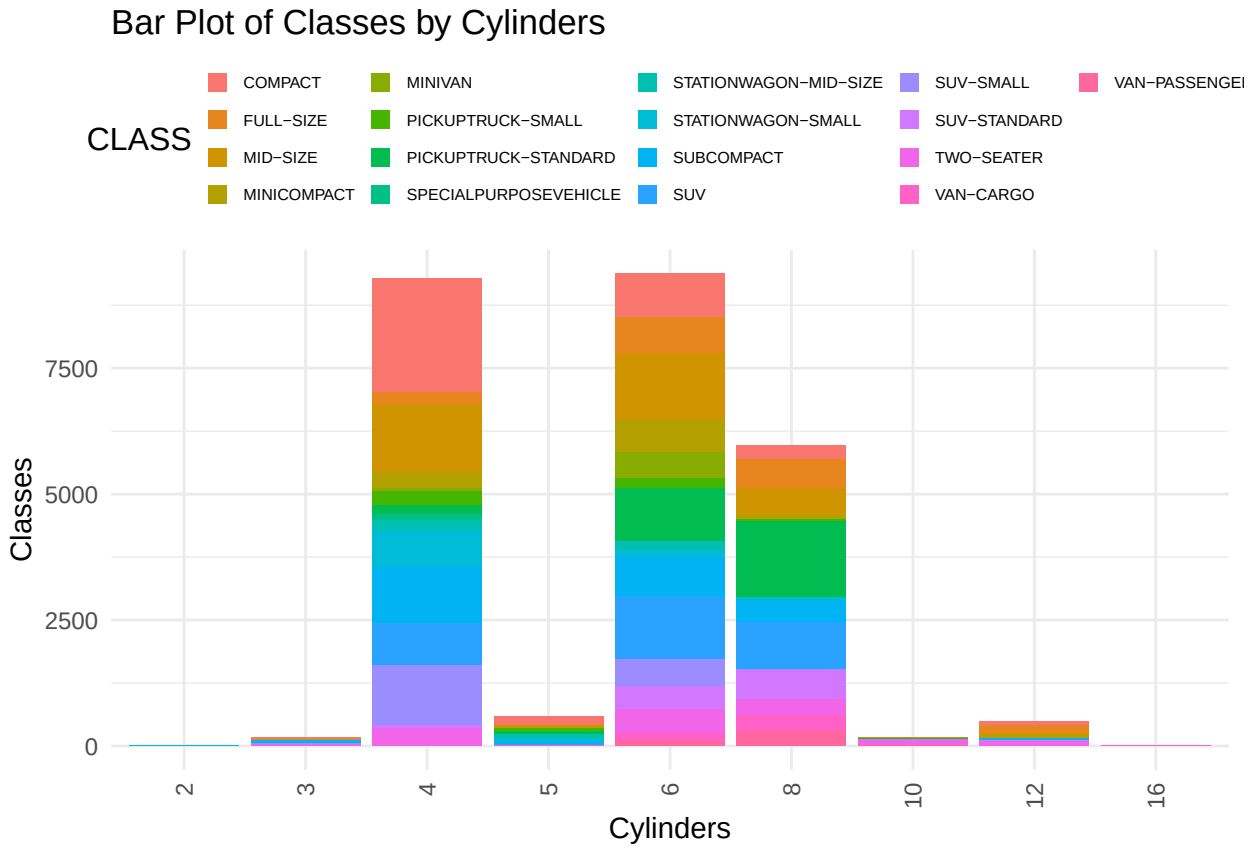
- There are 17 levels in CLASS, so with the purpose of analyze a way to consider this variable in the model, it was considered an alternative to unify categories and reduce the number of levels.


CLASS
COMPACT
FULL−SIZE
MID−SIZE
MINICOMPACT
MINIVAN
PICKUPTRUCK−SMALL
PICKUPTRUCK−STANDARD
SPECIALPURPOSEVEHICLE
STATIONWAGON−MID−SIZE
STATIONWAGON−SMALL
SUBCOMPACT
SUV
SUV−SMALL
SUV−STANDARD
TWO−SEATER
VAN−CARGO
VAN−PASSENGER
EMISSIONS
600
400
200
CLASS


CLASS
VAN−PASSENGER
VAN−CARGO
PICKUPTRUCK−STANDARD
SUV−STANDARD
SUV
PICKUPTRUCK−SMALL
FULL−SIZE
TWO−SEATER
MINIVAN
MINICOMPACT
STATIONWAGON−MID−SIZE
SPECIALPURPOSEVEHICLE
SUBCOMPACT
SUV−SMALL
MID−SIZE
COMPACT
STATIONWAGON−SMALL
EMISSIONS
600
400
200
CLASS

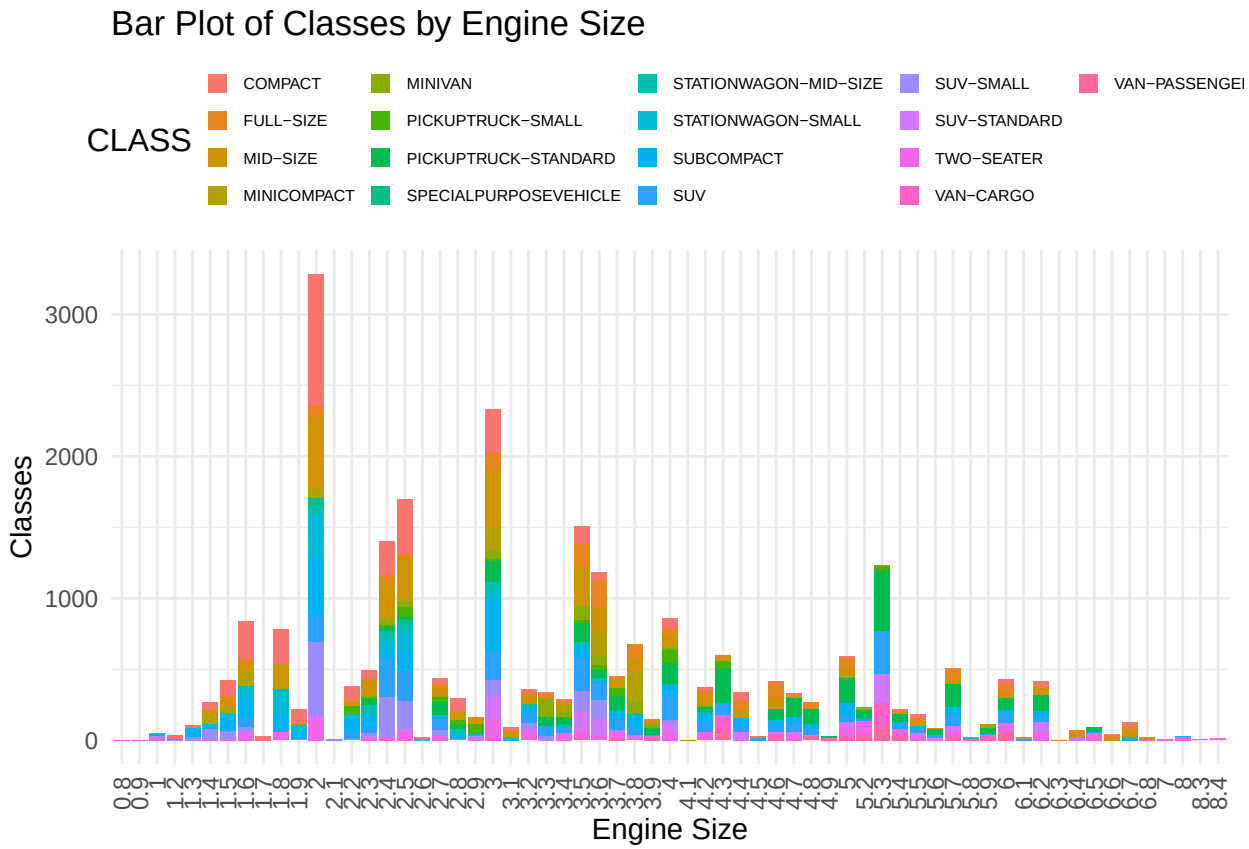
Insights

- In the case of unifying CLASS by the name, for example: Pickup Truck, Station Wagon, SUV, VAN, it would not be appropriate, since among these exist differences in the median emissions.



Insights

- There is no clear pattern if trying to unified classes by cylinders. There are many classes per each discrete number in the variable.



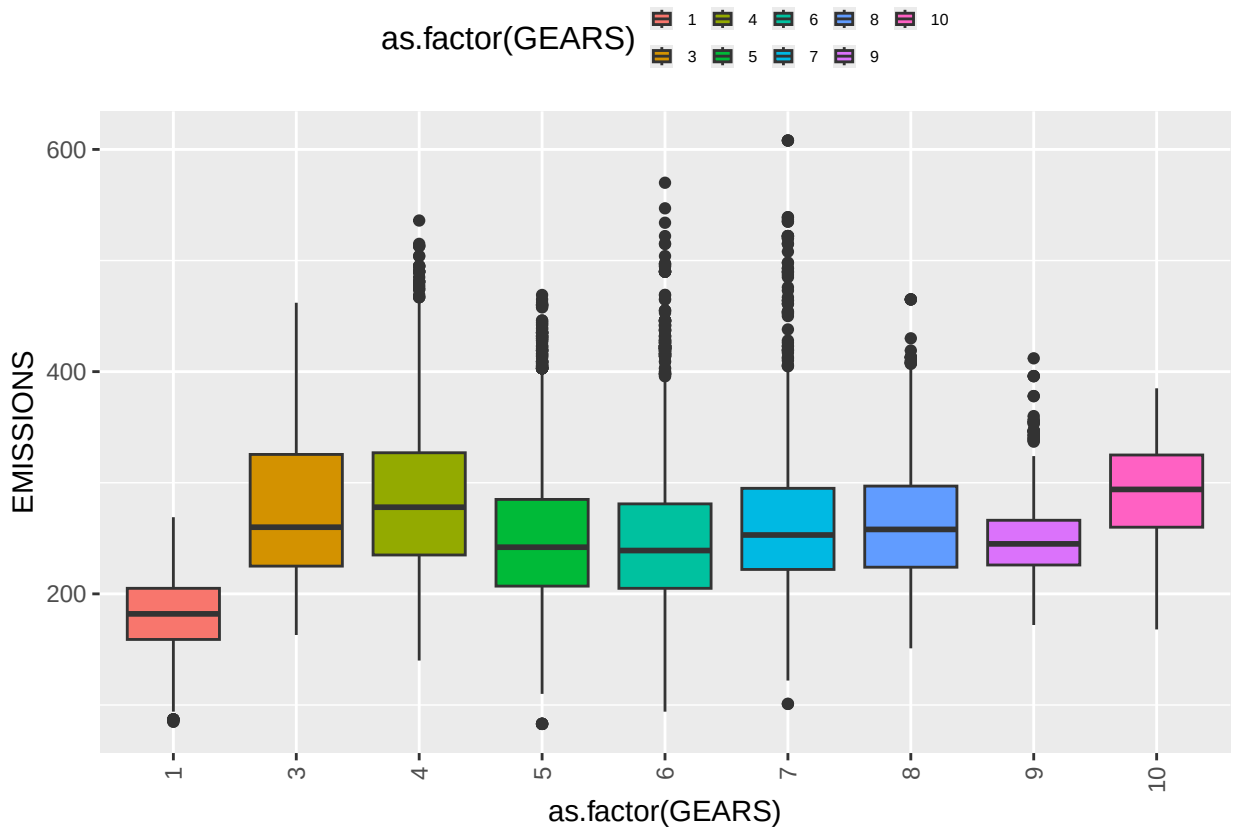
Insights

- There is no clear pattern if trying to unified classes by engine size.

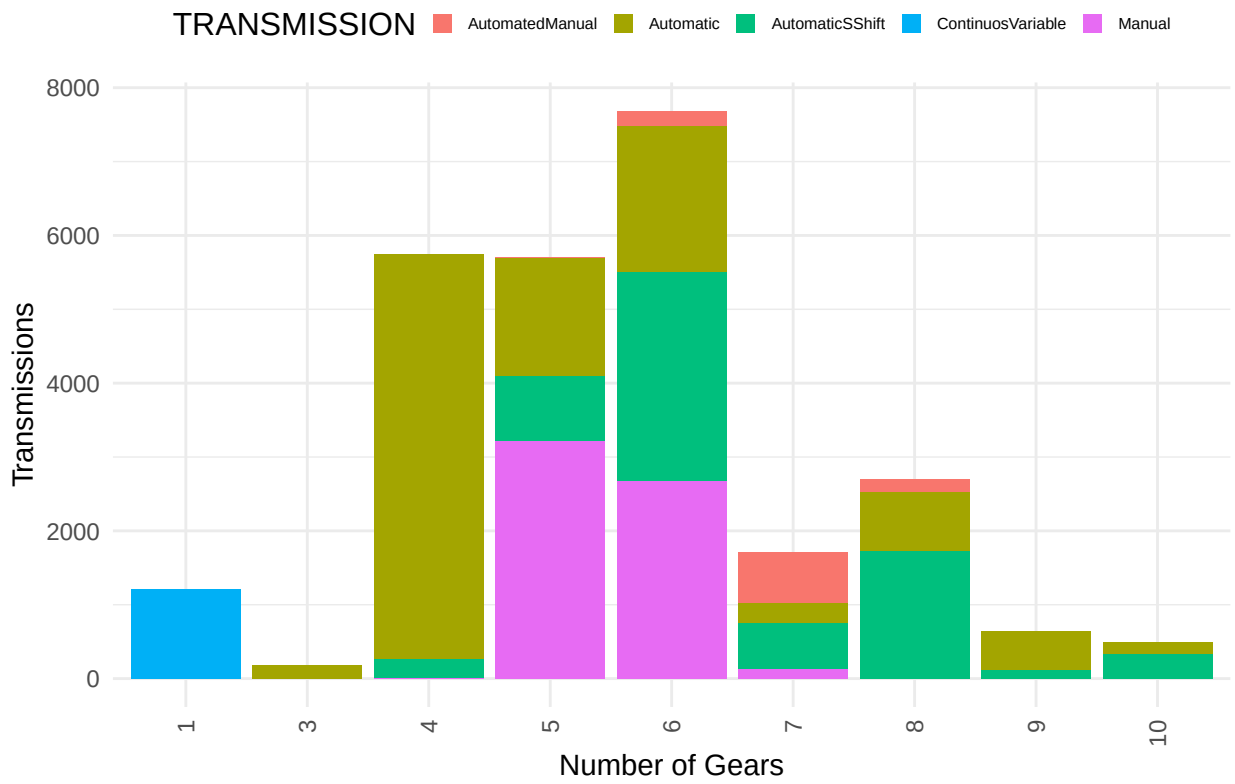
Conclusion

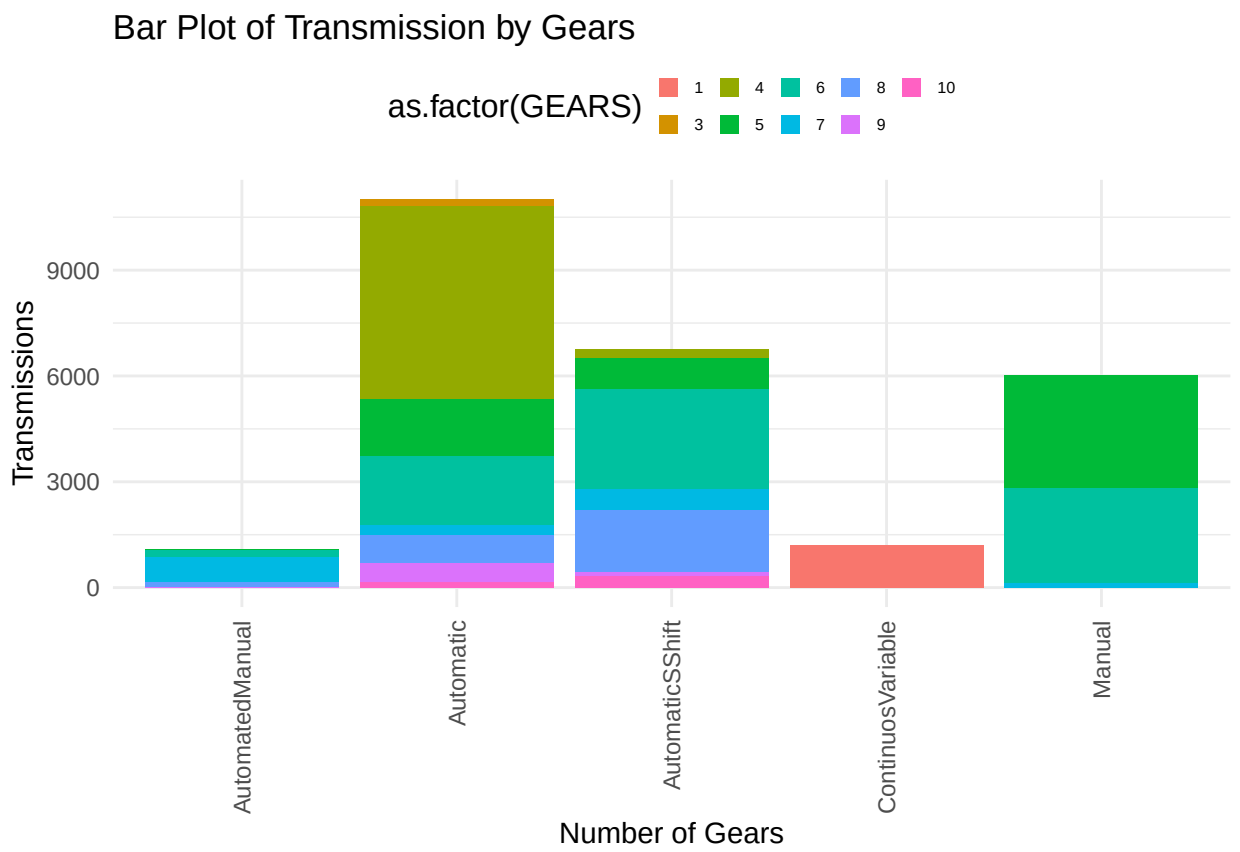
- Dropped the variable CLASS as a possible candidate to predictor in the model.

3.1.2.4 GEARS



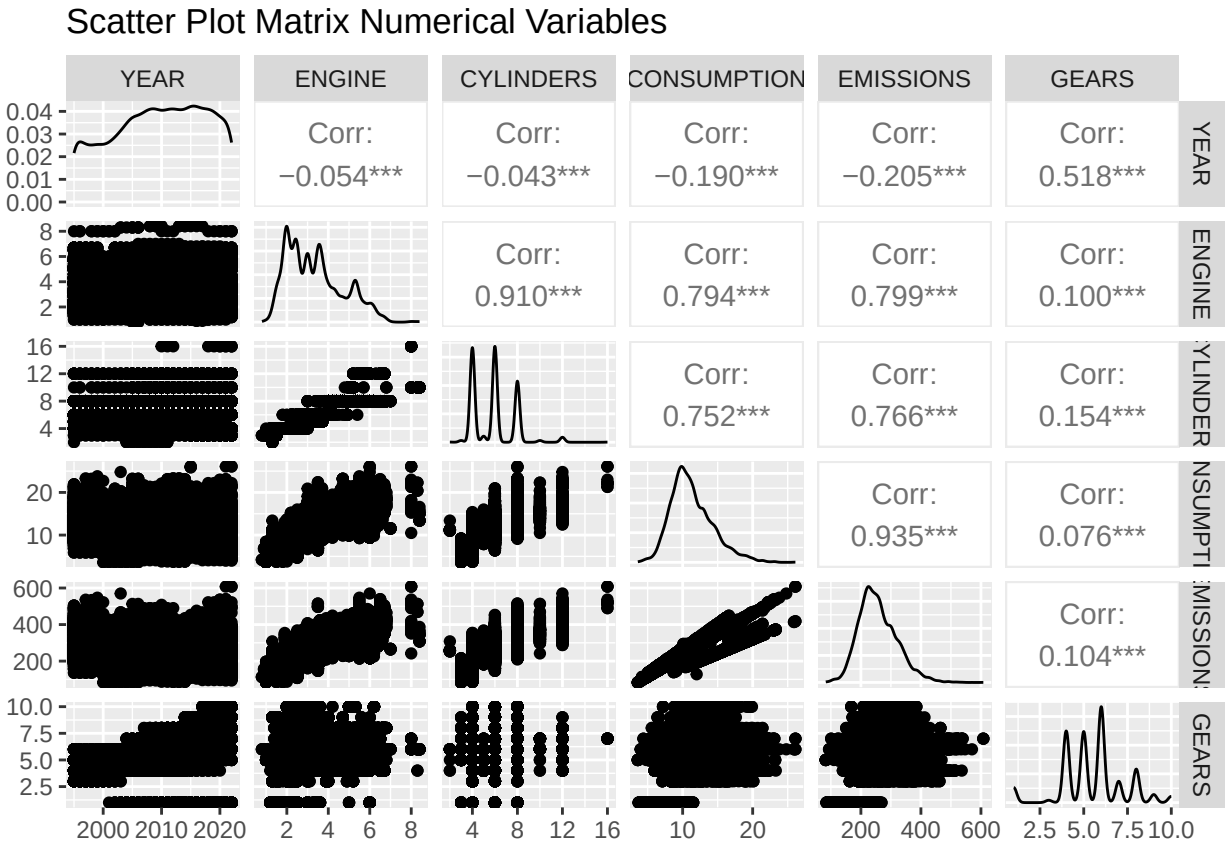
Bar Plot of Transmission by Gears





Insights

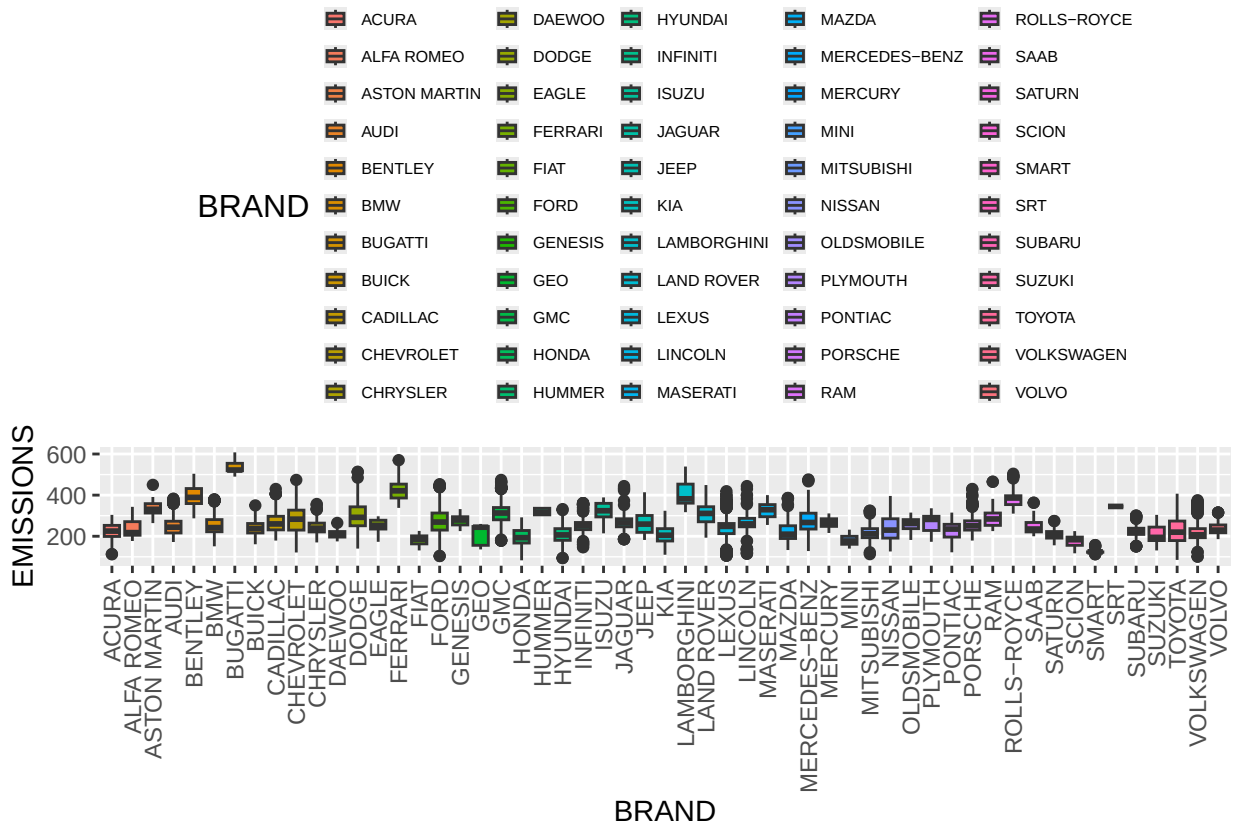
- Filled the NA values in Gears with 1 after analyzing that all the empty values corresponded to the Continuous Variable TRANSMISSION.
- For the purpose of consider GEARS in the model, it will be considered as numerical instead of categorical, and limit the number of coefficients (dummy variables).



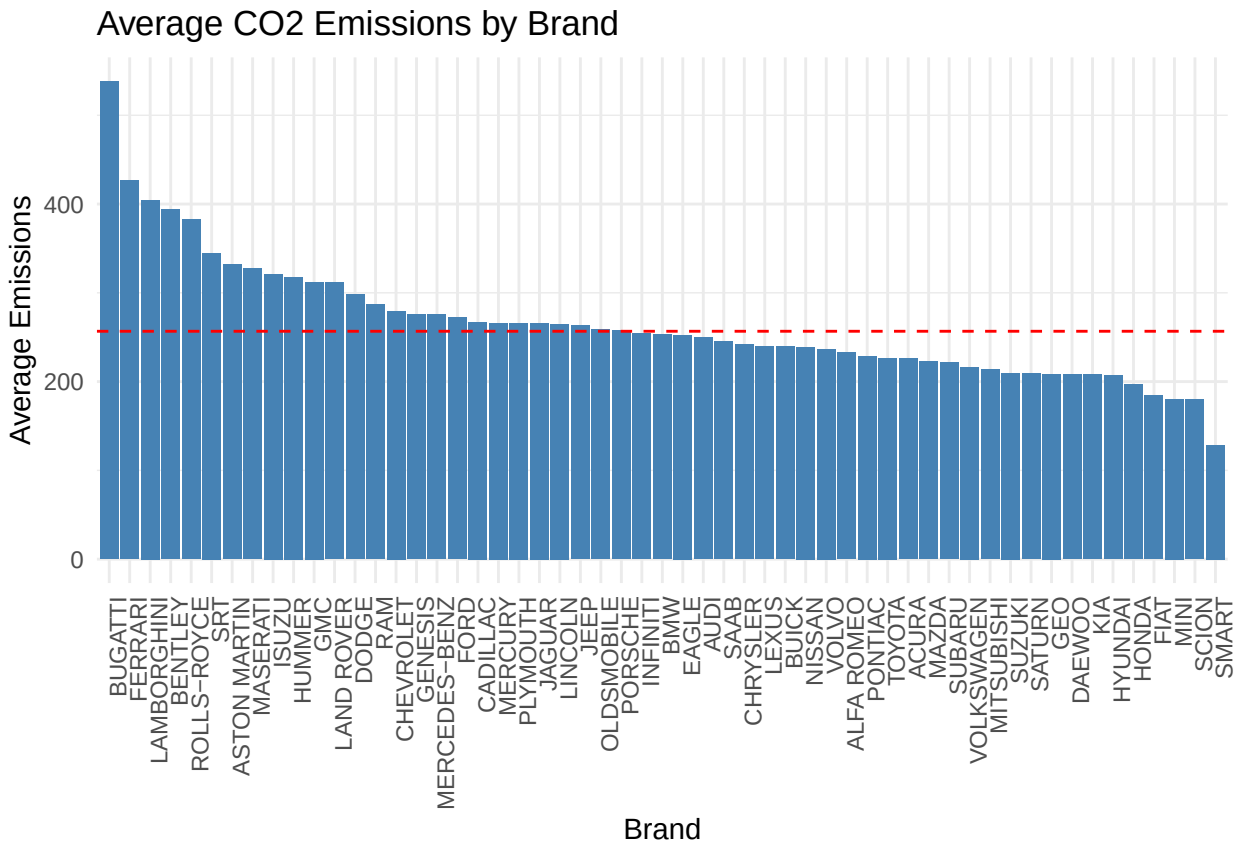
Insights

- GEARS is not greatly correlated with any of the other numerical variables.

3.1.2.5 BRAND




```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Insights

- The Brand with the highest average emission is Bugatti, and in general the brands that produce the high-end cars.
- Never the less, for its high amount of levels, dropped the variable for further analysis.

4 Data preparation

4.1 Training and Testing Data

```
set.seed(123)

test_index = sample(seq_len(nrow(co2_filter)),size=0.2*nrow(co2_filter))

co2_test = co2_filter[test_index, ]

co2_train = co2_filter[-test_index,]

write.csv(co2_train, "train.csv")

write.csv(co2_test, "test.csv")
```

```r
attach(co2_train)
```

5 Software model

5.1 Stepwise Regression

5.1.1 Both: Step

```r
step(lm(EMISSIONS~.,data = co2_train),direction = "both")
```

```
## Start:  AIC=64849.83
## EMISSIONS ~ YEAR + ENGINE + CYLINDERS + TRANSMISSION + GEARS +
##     FUEL + CONSUMPTION
##
##                Df Sum of Sq      RSS    AIC
## - CYLINDERS     1         1   466522  64848
## - ENGINE        1        31   466552  64849
## <none>                        466521  64850
## - TRANSMISSION  4      2476   468997  64952
## - YEAR          1      4811   471332  65062
## - GEARS         1      7135   473656  65164
## - FUEL          4   8533436  8999957 126581
## - CONSUMPTION   1  24657719 25124241 148002
##
## Step:  AIC=64847.86
## EMISSIONS ~ YEAR + ENGINE + TRANSMISSION + GEARS + FUEL + CONSUMPTION
##
##                Df Sum of Sq      RSS    AIC
## <none>                        466522  64848
## - ENGINE        1        59   466581  64849
## + CYLINDERS     1         1   466521  64850
## - TRANSMISSION  4      2505   469027  64952
## - YEAR          1      4823   471345  65060
## - GEARS         1      7172   473694  65164
## - FUEL          4   8591934  9058455 126714
## - CONSUMPTION   1  25194831 25661353 148441
##
## Call:
## lm(formula = EMISSIONS ~ YEAR + ENGINE + TRANSMISSION + GEARS +
##     FUEL + CONSUMPTION, data = co2_train)
##
## Coefficients:
##                   (Intercept)                           YEAR
##                     -170.72666                        0.10218
##                        ENGINE          TRANSMISSIONAutomatic
##                       -0.07035                       -0.63570
##   TRANSMISSIONAutomaticSShift  TRANSMISSIONContinuosVariable
##                       -0.69797                        1.55180
##            TRANSMISSIONManual                          GEARS
##                       -0.98620                        0.66221
##                     FUELEtanol                 FUELNaturalGas
##                     -151.11978                     -103.28606
##                    FUELPremium                    FUELRegular
```

```
##                     -35.14984                     -34.62523
##                   CONSUMPTION
##                      22.88428
```

5.1.2 Both: MASS

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

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

```
stepAIC(lm(EMISSIONS~.,data=co2_train),direction="both")
```

```
## Start:  AIC=64849.83
## EMISSIONS ~ YEAR + ENGINE + CYLINDERS + TRANSMISSION + GEARS +
##     FUEL + CONSUMPTION
##
##                Df Sum of Sq      RSS    AIC
## - CYLINDERS     1         1   466522  64848
## - ENGINE        1        31   466552  64849
## <none>                        466521  64850
## - TRANSMISSION  4      2476   468997  64952
## - YEAR          1      4811   471332  65062
## - GEARS         1      7135   473656  65164
## - FUEL          4   8533436  8999957 126581
## - CONSUMPTION   1  24657719 25124241 148002
##
## Step:  AIC=64847.86
## EMISSIONS ~ YEAR + ENGINE + TRANSMISSION + GEARS + FUEL + CONSUMPTION
##
##                Df Sum of Sq      RSS    AIC
## <none>                        466522  64848
## - ENGINE        1        59   466581  64849
## + CYLINDERS     1         1   466521  64850
## - TRANSMISSION  4      2505   469027  64952
## - YEAR          1      4823   471345  65060
## - GEARS         1      7172   473694  65164
## - FUEL          4   8591934  9058455 126714
## - CONSUMPTION   1  25194831 25661353 148441
```

```
##
## Call:
## lm(formula = EMISSIONS ~ YEAR + ENGINE + TRANSMISSION + GEARS +
##     FUEL + CONSUMPTION, data = co2_train)
##
## Coefficients:
##                   (Intercept)                           YEAR
##                     -170.72666                       0.10218
##                        ENGINE          TRANSMISSIONAutomatic
##                       -0.07035                      -0.63570
##   TRANSMISSIONAutomaticSShift  TRANSMISSIONContinuosVariable
##                       -0.69797                       1.55180
```

```
##            TRANSMISSIONManual                          GEARS
##                       -0.98620                        0.66221
##                     FUELEtanol                  FUELNaturalGas
##                     -151.11978                      -103.28606
##                    FUELPremium                     FUELRegular
##                      -35.14984                       -34.62523
##                    CONSUMPTION
##                       22.88428
```

5.1.3 Forward: List

```
mfl = lm(EMISSIONS~1,data=co2_train)
forwardAIC = step(mfl,scope=list(lower=~1,
                                 upper= ~ CONSUMPTION + CYLINDERS + ENGINE + FUEL + GEARS + TRANSMISSION
                  direction="forward", data=co2_train)
```

```
## Start:  AIC=172873.3
## EMISSIONS ~ 1
##
##                Df Sum of Sq      RSS    AIC
## + CONSUMPTION   1  72525464 10346035 129472
## + ENGINE        1  52862466 30009032 151686
## + CYLINDERS     1  48825735 34045763 154319
## + TRANSMISSION  4  11684751 71186747 169711
## + YEAR          1   3409758 79461741 171999
## + FUEL          4   1238568 81632930 172567
## + GEARS         1    930019 81941479 172640
## <none>                      82871498 172873
##
## Step:  AIC=129472.1
## EMISSIONS ~ CONSUMPTION
##
##                Df Sum of Sq      RSS    AIC
## + FUEL          4   9828471   517563  67000
## + CYLINDERS     1    766300  9579734 127869
## + ENGINE        1    698994  9647040 128015
## + GEARS         1     86144 10259891 129300
## + TRANSMISSION  4     85830 10260205 129306
## + YEAR          1     70951 10275083 129331
## <none>                      10346035 129472
##
## Step:  AIC=66999.68
## EMISSIONS ~ CONSUMPTION + FUEL
##
##                Df Sum of Sq    RSS   AIC
## + YEAR          1     41911 475653 65240
## + GEARS         1     31612 485952 65687
## + TRANSMISSION  4     11879 505685 66523
## + ENGINE        1       850 516714 66967
## + CYLINDERS     1       846 516718 66968
## <none>                      517563 67000
##
## Step:  AIC=65240.18
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR
```

```
##
##                Df Sum of Sq    RSS   AIC
## + GEARS         1    6597.5 469055 64951
## + TRANSMISSION  4    1886.5 473766 65165
## + ENGINE        1     149.7 475503 65236
## <none>                      475653 65240
## + CYLINDERS     1      15.5 475637 65241
##
## Step:  AIC=64950.81
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS
##
##                Df Sum of Sq    RSS   AIC
## + TRANSMISSION  4   2473.71 466581 64849
## <none>                      469055 64951
## + ENGINE        1     28.06 469027 64952
## + CYLINDERS     1      0.31 469055 64953
##
## Step:  AIC=64848.51
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION
##
##             Df Sum of Sq    RSS   AIC
## + ENGINE     1    59.419 466522 64848
## <none>                   466581 64849
## + CYLINDERS  1    29.058 466552 64849
##
## Step:  AIC=64847.86
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION +
##     ENGINE
##
##             Df Sum of Sq    RSS   AIC
## <none>                   466522 64848
## + CYLINDERS  1   0.62756 466521 64850
```

5.1.4 Forward: formula

```
mff=lm(EMISSIONS~CONSUMPTION + CYLINDERS + ENGINE + FUEL + GEARS + TRANSMISSION + YEAR)
forwardAIC1 <- step(mfl,scope=formula(mff),
                    direction="forward", data=co2_train)
```

```
## Start:  AIC=172873.3
## EMISSIONS ~ 1
##
##                Df Sum of Sq      RSS    AIC
## + CONSUMPTION   1  72525464 10346035 129472
## + ENGINE        1  52862466 30009032 151686
## + CYLINDERS     1  48825735 34045763 154319
## + TRANSMISSION  4  11684751 71186747 169711
## + YEAR          1   3409758 79461741 171999
## + FUEL          4   1238568 81632930 172567
## + GEARS         1    930019 81941479 172640
## <none>                      82871498 172873
##
## Step:  AIC=129472.1
## EMISSIONS ~ CONSUMPTION
```

```
##
##                Df Sum of Sq      RSS    AIC
## + FUEL          4   9828471   517563  67000
## + CYLINDERS     1    766300  9579734 127869
## + ENGINE        1    698994  9647040 128015
## + GEARS         1     86144 10259891 129300
## + TRANSMISSION  4     85830 10260205 129306
## + YEAR          1     70951 10275083 129331
## <none>                      10346035 129472
##
## Step:  AIC=66999.68
## EMISSIONS ~ CONSUMPTION + FUEL
##
##                Df Sum of Sq    RSS   AIC
## + YEAR          1     41911 475653 65240
## + GEARS         1     31612 485952 65687
## + TRANSMISSION  4     11879 505685 66523
## + ENGINE        1       850 516714 66967
## + CYLINDERS     1       846 516718 66968
## <none>                      517563 67000
##
## Step:  AIC=65240.18
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR
##
##                Df Sum of Sq    RSS   AIC
## + GEARS         1    6597.5 469055 64951
## + TRANSMISSION  4    1886.5 473766 65165
## + ENGINE        1     149.7 475503 65236
## <none>                      475653 65240
## + CYLINDERS     1      15.5 475637 65241
##
## Step:  AIC=64950.81
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS
##
##                Df Sum of Sq    RSS   AIC
## + TRANSMISSION  4   2473.71 466581 64849
## <none>                      469055 64951
## + ENGINE        1     28.06 469027 64952
## + CYLINDERS     1      0.31 469055 64953
##
## Step:  AIC=64848.51
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION
##
##             Df Sum of Sq    RSS   AIC
## + ENGINE     1    59.419 466522 64848
## <none>                   466581 64849
## + CYLINDERS  1    29.058 466552 64849
##
## Step:  AIC=64847.86
## EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION +
##     ENGINE
##
##             Df Sum of Sq    RSS   AIC
## <none>                   466522 64848
```

```
## + CYLINDERS  1   0.62756 466521 64850
```

```
forwardAIC1$coefficients
```

```
##                   (Intercept)                   CONSUMPTION
##                 -170.72665847                   22.88427986
##                    FUELEtanol                FUELNaturalGas
##                 -151.11978335                -103.28606140
##                   FUELPremium                   FUELRegular
##                  -35.14983959                 -34.62523342
##                          YEAR                         GEARS
##                    0.10218330                   0.66220824
##         TRANSMISSIONAutomatic   TRANSMISSIONAutomaticSShift
##                    -0.63570067                  -0.69797284
## TRANSMISSIONContinuosVariable            TRANSMISSIONManual
##                     1.55179737                  -0.98620435
##                         ENGINE
##                    -0.07034717
```

```
forwardAIC1$anova
```

```
##              Step Df     Deviance Resid. Df Resid. Dev       AIC
## 1                 NA           NA     20859 82871498.1 172873.26
## 2  + CONSUMPTION -1 7.252546e+07     20858 10346034.6 129472.11
## 3         + FUEL -4 9.828471e+06     20854   517563.3  66999.68
## 4         + YEAR -1 4.191075e+04     20853   475652.6  65240.18
## 5        + GEARS -1 6.597537e+03     20852   469055.0  64950.81
## 6 + TRANSMISSION -4 2.473711e+03     20848   466581.3  64848.51
## 7       + ENGINE -1 5.941885e+01     20847   466521.9  64847.86
```

5.1.5 Backward: Step

```
step(lm(EMISSIONS~.,data = co2_train),direction = "backward")
```

```
## Start:  AIC=64849.83
## EMISSIONS ~ YEAR + ENGINE + CYLINDERS + TRANSMISSION + GEARS +
##     FUEL + CONSUMPTION
##
##                Df Sum of Sq      RSS    AIC
## - CYLINDERS     1         1   466522  64848
## - ENGINE        1        31   466552  64849
## <none>                       466521  64850
## - TRANSMISSION  4      2476   468997  64952
## - YEAR          1      4811   471332  65062
## - GEARS         1      7135   473656  65164
## - FUEL          4   8533436  8999957 126581
## - CONSUMPTION   1  24657719 25124241 148002
##
## Step:  AIC=64847.86
## EMISSIONS ~ YEAR + ENGINE + TRANSMISSION + GEARS + FUEL + CONSUMPTION
##
##                Df Sum of Sq      RSS    AIC
## <none>                       466522  64848
## - ENGINE        1        59   466581  64849
## - TRANSMISSION  4      2505   469027  64952
## - YEAR          1      4823   471345  65060
```

```
## - GEARS         1      7172   473694  65164
## - FUEL          4   8591934  9058455 126714
## - CONSUMPTION   1  25194831 25661353 148441

##
## Call:
## lm(formula = EMISSIONS ~ YEAR + ENGINE + TRANSMISSION + GEARS +
##     FUEL + CONSUMPTION, data = co2_train)
##
## Coefficients:
##                   (Intercept)                           YEAR
##                    -170.72666                        0.10218
##                        ENGINE          TRANSMISSIONAutomatic
##                      -0.07035                       -0.63570
##   TRANSMISSIONAutomaticSShift  TRANSMISSIONContinuosVariable
##                      -0.69797                        1.55180
##            TRANSMISSIONManual                          GEARS
##                      -0.98620                        0.66221
##                    FUELEtanol                 FUELNaturalGas
##                    -151.11978                     -103.28606
##                   FUELPremium                    FUELRegular
##                     -35.14984                      -34.62523
##                   CONSUMPTION
##                      22.88428
```

5.2 All Possible Regression Subset

5.2.1 Regsubset

```
library(leaps)
subset_model1 = regsubsets(EMISSIONS~CONSUMPTION + CYLINDERS + ENGINE + FUEL + GEARS + TRANSMISSION + YE
summary(subset_model1)
```

```
## Subset selection object
## Call: regsubsets.formula(EMISSIONS ~ CONSUMPTION + CYLINDERS + ENGINE +
##     FUEL + GEARS + TRANSMISSION + YEAR, data = co2_train, nvmax = 15)
## 13 Variables  (and intercept)
##                               Forced in Forced out
## CONSUMPTION                       FALSE      FALSE
## CYLINDERS                         FALSE      FALSE
## ENGINE                            FALSE      FALSE
## FUELEtanol                        FALSE      FALSE
## FUELNaturalGas                    FALSE      FALSE
## FUELPremium                       FALSE      FALSE
## FUELRegular                       FALSE      FALSE
## GEARS                             FALSE      FALSE
## TRANSMISSIONAutomatic             FALSE      FALSE
## TRANSMISSIONAutomaticSShift       FALSE      FALSE
## TRANSMISSIONContinuosVariable     FALSE      FALSE
## TRANSMISSIONManual                FALSE      FALSE
## YEAR                              FALSE      FALSE
## 1 subsets of each size up to 13
## Selection Algorithm: exhaustive
##          CONSUMPTION CYLINDERS ENGINE FUELEtanol FUELNaturalGas FUELPremium
## 1  ( 1 )  "*"         " "       " "    " "         " "             " "
```

```
## 2  ( 1 )  "*"         " "        " "    "*"        " "            " "
## 3  ( 1 )  "*"         " "        " "    "*"        "*"            " "
## 4  ( 1 )  "*"         " "        " "    "*"        " "            "*"
## 5  ( 1 )  "*"         " "        " "    "*"        "*"            "*"
## 6  ( 1 )  "*"         " "        " "    "*"        "*"            "*"
## 7  ( 1 )  "*"         " "        " "    "*"        "*"            "*"
## 8  ( 1 )  "*"         " "        " "    "*"        "*"            "*"
## 9  ( 1 )  "*"         " "        " "    "*"        "*"            "*"
## 10  ( 1 ) "*"         " "        "*"    "*"        "*"            "*"
## 11  ( 1 ) "*"         " "        " "    "*"        "*"            "*"
## 12  ( 1 ) "*"         " "        "*"    "*"        "*"            "*"
## 13  ( 1 ) "*"         "*"        "*"    "*"        "*"            "*"
##           FUELRegular GEARS TRANSMISSIONAutomatic TRANSMISSIONAutomaticSShift
## 1  ( 1 )  " "         " "    " "                  " "
## 2  ( 1 )  " "         " "    " "                  " "
## 3  ( 1 )  " "         " "    " "                  " "
## 4  ( 1 )  "*"         " "    " "                  " "
## 5  ( 1 )  "*"         " "    " "                  " "
## 6  ( 1 )  "*"         " "    " "                  " "
## 7  ( 1 )  "*"         "*"    " "                  " "
## 8  ( 1 )  "*"         "*"    " "                  " "
## 9  ( 1 )  "*"         "*"    " "                  " "
## 10  ( 1 ) "*"         "*"    " "                  " "
## 11  ( 1 ) "*"         "*"    "*"                  "*"
## 12  ( 1 ) "*"         "*"    "*"                  "*"
## 13  ( 1 ) "*"         "*"    "*"                  "*"
##           TRANSMISSIONContinuosVariable TRANSMISSIONManual YEAR
## 1  ( 1 )  " "                           " "                " "
## 2  ( 1 )  " "                           " "                " "
## 3  ( 1 )  " "                           " "                " "
## 4  ( 1 )  " "                           " "                " "
## 5  ( 1 )  " "                           " "                " "
## 6  ( 1 )  " "                           " "                "*"
## 7  ( 1 )  " "                           " "                "*"
## 8  ( 1 )  "*"                           " "                "*"
## 9  ( 1 )  "*"                           "*"                "*"
## 10  ( 1 ) "*"                           "*"                "*"
## 11  ( 1 ) "*"                           "*"                "*"
## 12  ( 1 ) "*"                           "*"                "*"
## 13  ( 1 ) "*"                           "*"                "*"
```

5.2.2 Results

```
SUM = summary(subset_model1)
names(SUM)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

```
Rsq=SUM$rsq
CP=SUM$cp
AdRsq=SUM$adjr2
BIC=SUM$bic
RSS=SUM$rss

cbind(SUM$which,round(cbind(Rsq,AdRsq,CP,BIC,RSS),4))
```

```
##    (Intercept) CONSUMPTION CYLINDERS ENGINE FUELEtanol FUELNaturalGas
## 1            1           1         0      0          0              0
## 2            1           1         0      0          1              0
## 3            1           1         0      0          1              1
## 4            1           1         0      0          1              0
## 5            1           1         0      0          1              1
## 6            1           1         0      0          1              1
## 7            1           1         0      0          1              1
## 8            1           1         0      0          1              1
## 9            1           1         0      0          1              1
## 10           1           1         0      1          1              1
## 11           1           1         0      0          1              1
## 12           1           1         0      1          1              1
## 13           1           1         1      1          1              1
##    FUELPremium FUELRegular GEARS TRANSMISSIONAutomatic
## 1            0           0     0                     0
## 2            0           0     0                     0
## 3            0           0     0                     0
## 4            1           1     0                     0
## 5            1           1     0                     0
## 6            1           1     0                     0
## 7            1           1     1                     0
## 8            1           1     1                     0
## 9            1           1     1                     0
## 10           1           1     1                     0
## 11           1           1     1                     1
## 12           1           1     1                     1
## 13           1           1     1                     1
##    TRANSMISSIONAutomaticSShift TRANSMISSIONContinuosVariable TRANSMISSIONManual
## 1                            0                             0                  0
## 2                            0                             0                  0
## 3                            0                             0                  0
## 4                            0                             0                  0
## 5                            0                             0                  0
## 6                            0                             0                  0
## 7                            0                             0                  0
## 8                            0                             1                  0
## 9                            0                             1                  1
## 10                           0                             1                  1
## 11                           1                             1                  1
## 12                           1                             1                  1
## 13                           1                             1                  1
##    YEAR    Rsq  AdRsq          CP        BIC        RSS
## 1     0 0.8752 0.8751 441445.4143  -43383.26 10346034.6
## 2     0 0.9874 0.9874  25744.4960  -91239.69  1042847.0
## 3     0 0.9889 0.9889  20160.8681  -93893.20   917843.9
## 4     0 0.9906 0.9906  14023.4985  -97265.88   780448.4
## 5     0 0.9938 0.9938   2278.7587 -105823.91   517563.3
## 6     1 0.9943 0.9943    408.0221 -107575.47   475652.6
## 7     1 0.9943 0.9943    115.2182 -107856.88   469055.0
## 8     1 0.9944 0.9944     41.7123 -107922.22   467365.2
## 9     1 0.9944 0.9944     25.6971 -107930.28   466962.1
```

```
## 10    1 0.9944 0.9944     23.9186 -107924.11   466877.5
## 11    1 0.9944 0.9944     12.6831 -107927.40   466581.3
## 12    1 0.9944 0.9944     12.0280 -107920.12   466521.9
## 13    1 0.9944 0.9944     14.0000 -107910.20   466521.3
```

```
n = length(co2_train$EMISSIONS) #number of observations
p = apply(SUM$which, 1, sum) #number of variables
```

```
#Calculation of AIC
AIC = SUM$bic - log(n) * p + 2 * p

#number of independent variables in the models
I=p-1
MSE1=RSS/(n-I-1)
```

```
SUM$which
```

```
##    (Intercept) CONSUMPTION CYLINDERS ENGINE FUELEtanol FUELNaturalGas
## 1         TRUE        TRUE     FALSE  FALSE      FALSE          FALSE
## 2         TRUE        TRUE     FALSE  FALSE       TRUE          FALSE
## 3         TRUE        TRUE     FALSE  FALSE       TRUE           TRUE
## 4         TRUE        TRUE     FALSE  FALSE       TRUE          FALSE
## 5         TRUE        TRUE     FALSE  FALSE       TRUE           TRUE
## 6         TRUE        TRUE     FALSE  FALSE       TRUE           TRUE
## 7         TRUE        TRUE     FALSE  FALSE       TRUE           TRUE
## 8         TRUE        TRUE     FALSE  FALSE       TRUE           TRUE
## 9         TRUE        TRUE     FALSE  FALSE       TRUE           TRUE
## 10        TRUE        TRUE     FALSE   TRUE       TRUE           TRUE
## 11        TRUE        TRUE     FALSE  FALSE       TRUE           TRUE
## 12        TRUE        TRUE     FALSE   TRUE       TRUE           TRUE
## 13        TRUE        TRUE      TRUE   TRUE       TRUE           TRUE
##    FUELPremium FUELRegular GEARS TRANSMISSIONAutomatic
## 1        FALSE       FALSE FALSE                 FALSE
## 2        FALSE       FALSE FALSE                 FALSE
## 3        FALSE       FALSE FALSE                 FALSE
## 4         TRUE        TRUE FALSE                 FALSE
## 5         TRUE        TRUE FALSE                 FALSE
## 6         TRUE        TRUE FALSE                 FALSE
## 7         TRUE        TRUE  TRUE                 FALSE
## 8         TRUE        TRUE  TRUE                 FALSE
## 9         TRUE        TRUE  TRUE                 FALSE
## 10        TRUE        TRUE  TRUE                 FALSE
## 11        TRUE        TRUE  TRUE                  TRUE
## 12        TRUE        TRUE  TRUE                  TRUE
## 13        TRUE        TRUE  TRUE                  TRUE
##    TRANSMISSIONAutomaticSShift TRANSMISSIONContinuosVariable TRANSMISSIONManual
## 1                        FALSE                         FALSE              FALSE
## 2                        FALSE                         FALSE              FALSE
## 3                        FALSE                         FALSE              FALSE
## 4                        FALSE                         FALSE              FALSE
## 5                        FALSE                         FALSE              FALSE
## 6                        FALSE                         FALSE              FALSE
## 7                        FALSE                         FALSE              FALSE
## 8                        FALSE                          TRUE              FALSE
## 9                        FALSE                          TRUE               TRUE
```

```
## 10                        FALSE                           TRUE               TRUE
## 11                         TRUE                           TRUE               TRUE
## 12                         TRUE                           TRUE               TRUE
## 13                         TRUE                           TRUE               TRUE
##     YEAR
## 1  FALSE
## 2  FALSE
## 3  FALSE
## 4  FALSE
## 5  FALSE
## 6   TRUE
## 7   TRUE
## 8   TRUE
## 9   TRUE
## 10  TRUE
## 11  TRUE
## 12  TRUE
## 13  TRUE
```

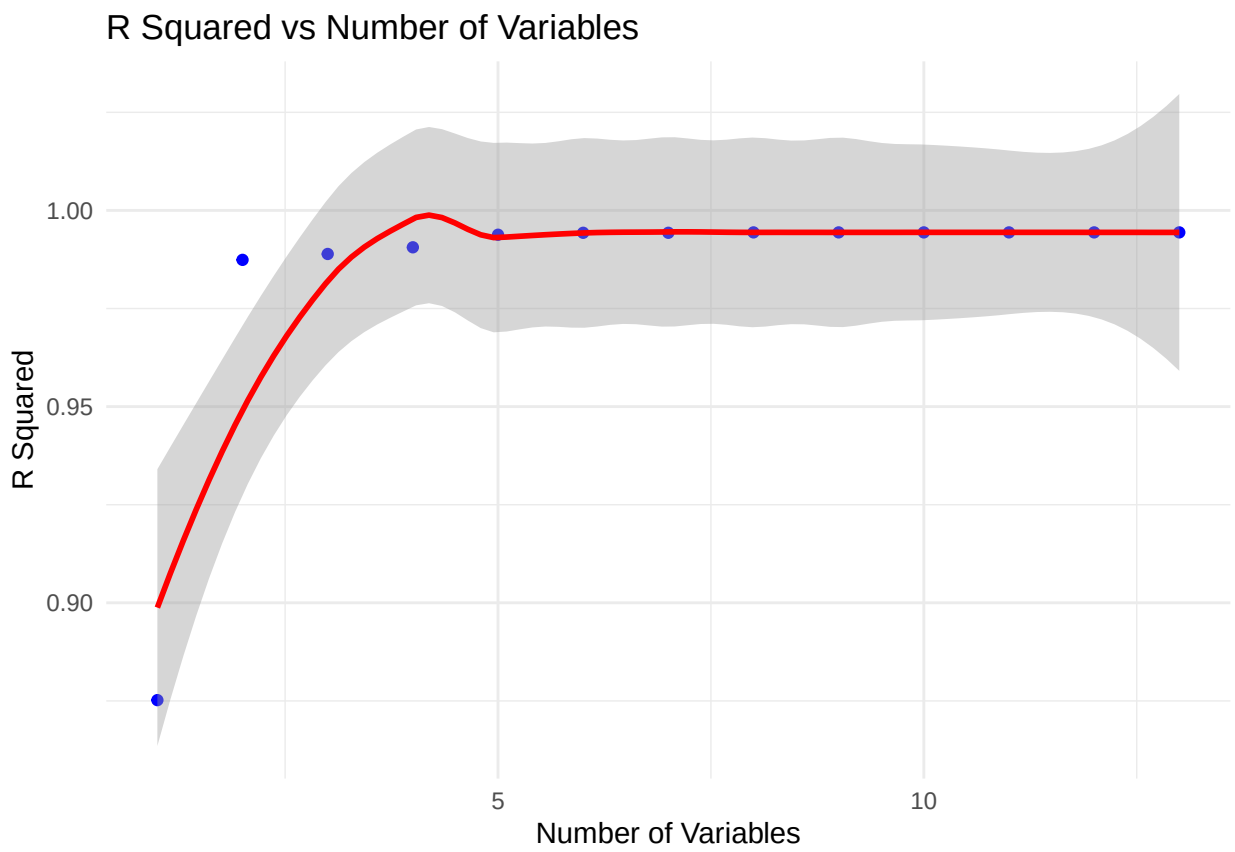
```
results_df = data.frame(
  Model = SUM$which,
  Rsq = round(Rsq, 4),
  AdRsq = round(AdRsq, 4),
  CP = round(CP, 4),
  BIC = round(BIC, 4),
  RSS = round(RSS, 4),
  AIC = round(AIC, 4),
  MSE1 = round(MSE1, 4)
)

results_df
```

```
##    Model..Intercept. Model.CONSUMPTION Model.CYLINDERS Model.ENGINE
## 1               TRUE              TRUE           FALSE        FALSE
## 2               TRUE              TRUE           FALSE        FALSE
## 3               TRUE              TRUE           FALSE        FALSE
## 4               TRUE              TRUE           FALSE        FALSE
## 5               TRUE              TRUE           FALSE        FALSE
## 6               TRUE              TRUE           FALSE        FALSE
## 7               TRUE              TRUE           FALSE        FALSE
## 8               TRUE              TRUE           FALSE        FALSE
## 9               TRUE              TRUE           FALSE        FALSE
## 10              TRUE              TRUE           FALSE         TRUE
## 11              TRUE              TRUE           FALSE        FALSE
## 12              TRUE              TRUE           FALSE         TRUE
## 13              TRUE              TRUE            TRUE         TRUE
##    Model.FUELEtanol Model.FUELNaturalGas Model.FUELPremium Model.FUELRegular
## 1             FALSE                FALSE             FALSE             FALSE
## 2              TRUE                FALSE             FALSE             FALSE
## 3              TRUE                 TRUE             FALSE             FALSE
## 4              TRUE                FALSE              TRUE              TRUE
## 5              TRUE                 TRUE              TRUE              TRUE
## 6              TRUE                 TRUE              TRUE              TRUE
## 7              TRUE                 TRUE              TRUE              TRUE
## 8              TRUE                 TRUE              TRUE              TRUE
```

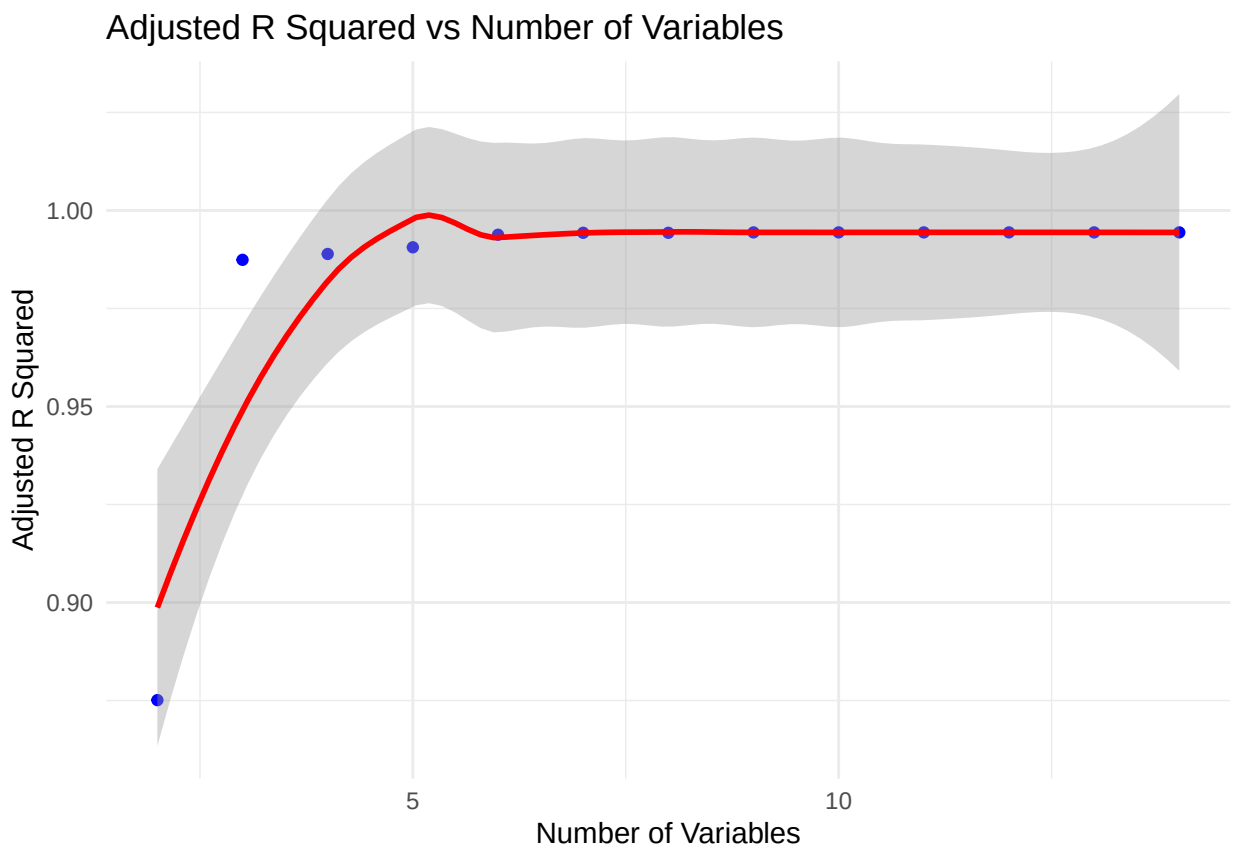
```
## 9              TRUE                  TRUE              TRUE              TRUE
## 10             TRUE                  TRUE              TRUE              TRUE
## 11             TRUE                  TRUE              TRUE              TRUE
## 12             TRUE                  TRUE              TRUE              TRUE
## 13             TRUE                  TRUE              TRUE              TRUE
##    Model.GEARS Model.TRANSMISSIONAutomatic Model.TRANSMISSIONAutomaticSShift
## 1        FALSE                       FALSE                             FALSE
## 2        FALSE                       FALSE                             FALSE
## 3        FALSE                       FALSE                             FALSE
## 4        FALSE                       FALSE                             FALSE
## 5        FALSE                       FALSE                             FALSE
## 6        FALSE                       FALSE                             FALSE
## 7         TRUE                       FALSE                             FALSE
## 8         TRUE                       FALSE                             FALSE
## 9         TRUE                       FALSE                             FALSE
## 10        TRUE                       FALSE                             FALSE
## 11        TRUE                        TRUE                              TRUE
## 12        TRUE                        TRUE                              TRUE
## 13        TRUE                        TRUE                              TRUE
##    Model.TRANSMISSIONContinuosVariable Model.TRANSMISSIONManual Model.YEAR
## 1                                FALSE                    FALSE      FALSE
## 2                                FALSE                    FALSE      FALSE
## 3                                FALSE                    FALSE      FALSE
## 4                                FALSE                    FALSE      FALSE
## 5                                FALSE                    FALSE      FALSE
## 6                                FALSE                    FALSE       TRUE
## 7                                FALSE                    FALSE       TRUE
## 8                                 TRUE                    FALSE       TRUE
## 9                                 TRUE                     TRUE       TRUE
## 10                                TRUE                     TRUE       TRUE
## 11                                TRUE                     TRUE       TRUE
## 12                                TRUE                     TRUE       TRUE
## 13                                TRUE                     TRUE       TRUE
##       Rsq  AdRsq           CP        BIC        RSS        AIC     MSE1
## 1  0.8752 0.8751 441445.4143  -43383.26 10346034.6  -43399.15 496.0224
## 2  0.9874 0.9874  25744.4960  -91239.69  1042847.0  -91263.52  49.9999
## 3  0.9889 0.9889  20160.8681  -93893.20   917843.9  -93924.98  44.0086
## 4  0.9906 0.9906  14023.4985  -97265.88   780448.4  -97305.61  37.4226
## 5  0.9938 0.9938   2278.7587 -105823.91   517563.3 -105871.58  24.8184
## 6  0.9943 0.9943    408.0221 -107575.47   475652.6 -107631.09  22.8098
## 7  0.9943 0.9943    115.2182 -107856.88   469055.0 -107920.45  22.4945
## 8  0.9944 0.9944     41.7123 -107922.22   467365.2 -107993.73  22.4145
## 9  0.9944 0.9944     25.6971 -107930.28   466962.1 -108009.74  22.3963
## 10 0.9944 0.9944     23.9186 -107924.11   466877.5 -108011.51  22.3933
## 11 0.9944 0.9944     12.6831 -107927.40   466581.3 -108022.75  22.3801
## 12 0.9944 0.9944     12.0280 -107920.12   466521.9 -108023.41  22.3784
## 13 0.9944 0.9944     14.0000 -107910.20   466521.3 -108021.44  22.3794
```

5.2.3 Plots

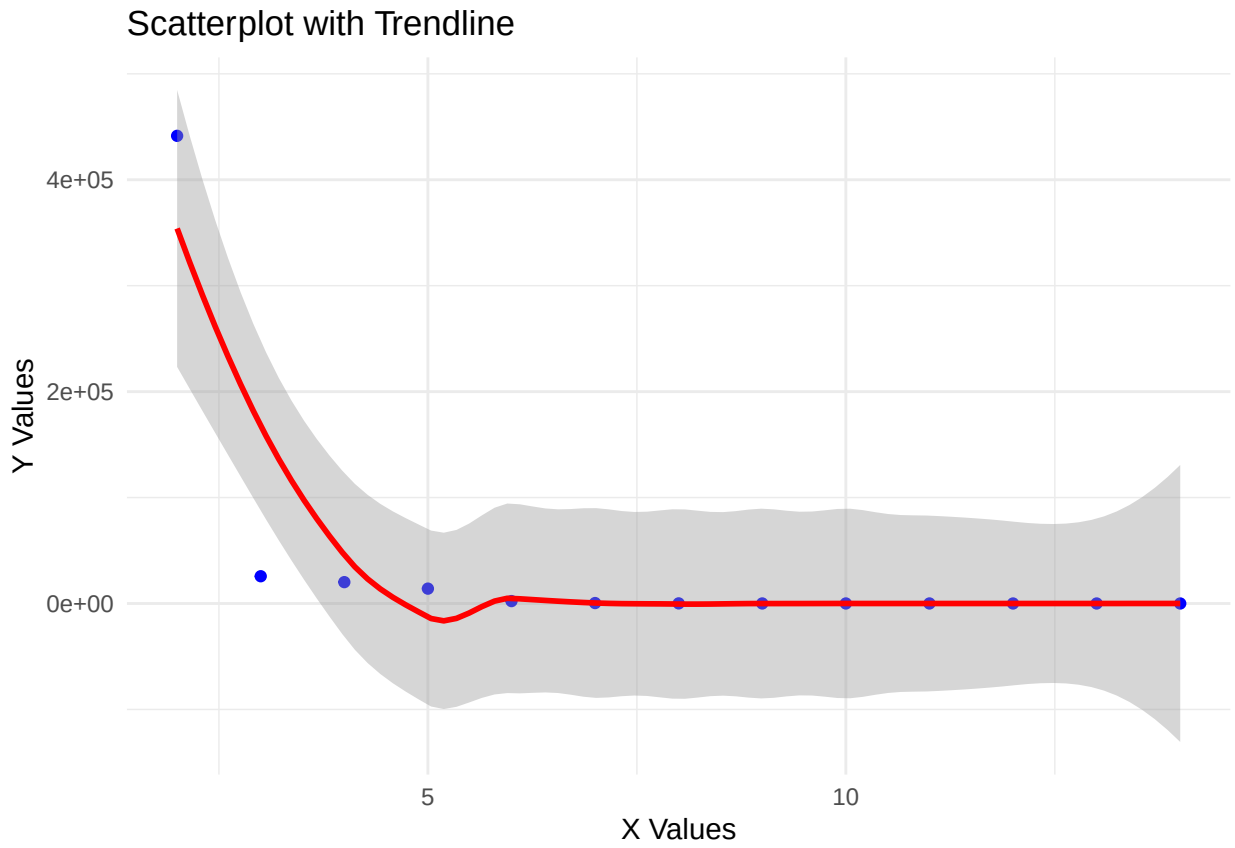
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```



Insights

- All the screening methods had the same result:

EMISSIONS ~ CONSUMPTION+FUEL+YEAR+GEARS+TRANSMISSION+ENGINE

- The adjusted r-squared was very high for all the options reviewed in the all possible method, which could lead to considering smaller models.

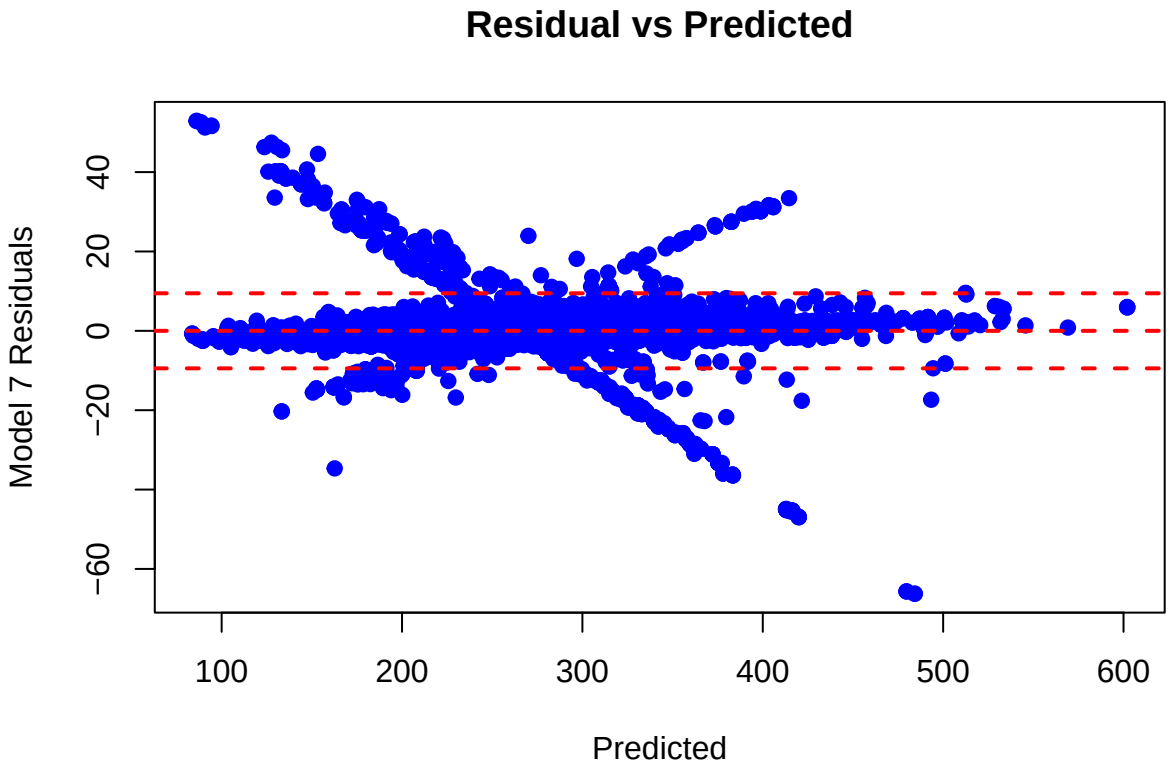
5.2.4 Model 7: result from screening methods

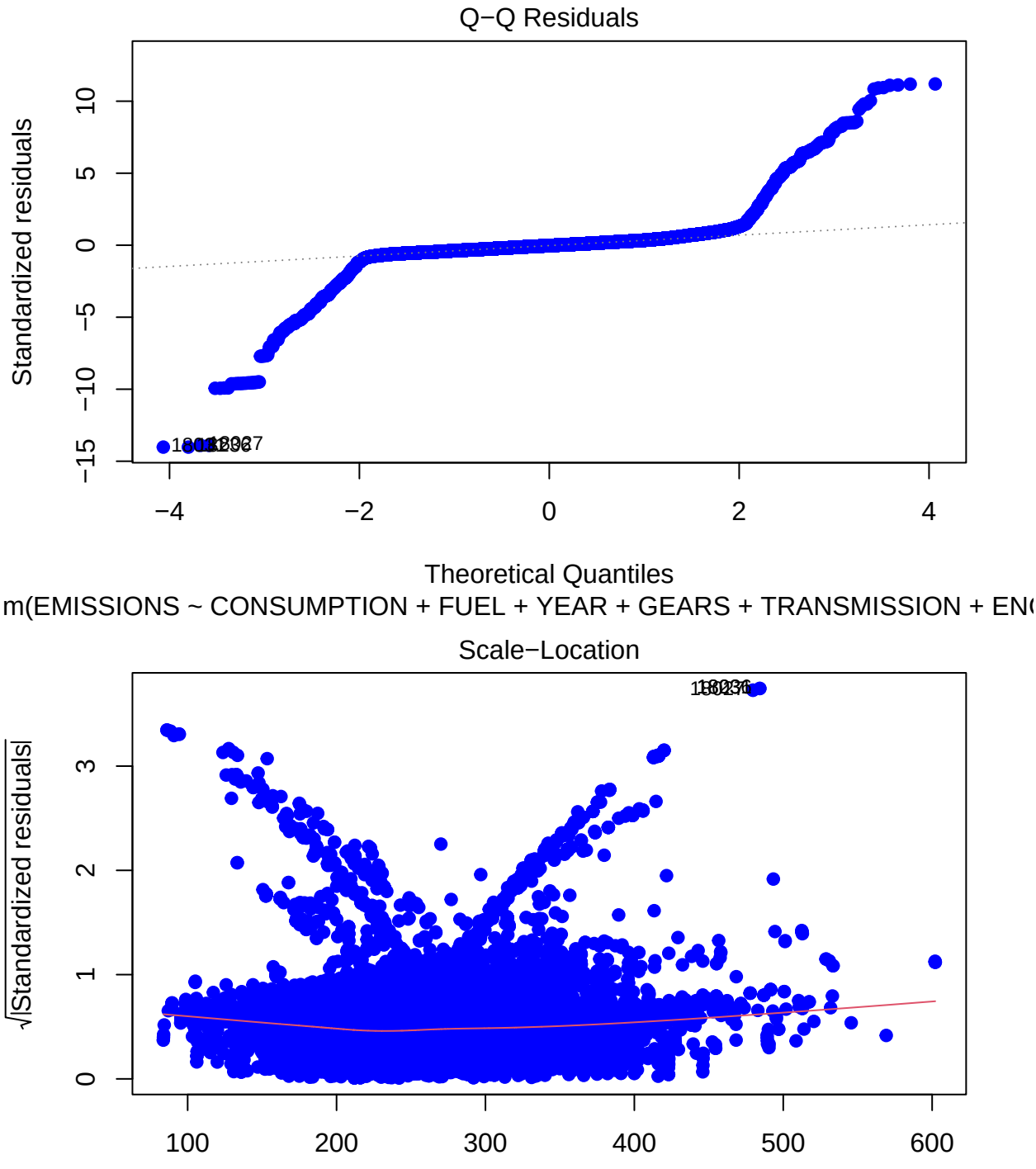
```
model7 = lm(data = co2_train,
            formula = EMISSIONS ~ CONSUMPTION+FUEL+YEAR+GEARS+TRANSMISSION+ENGINE)
s_model7 = summary(model7)
s_model7
```

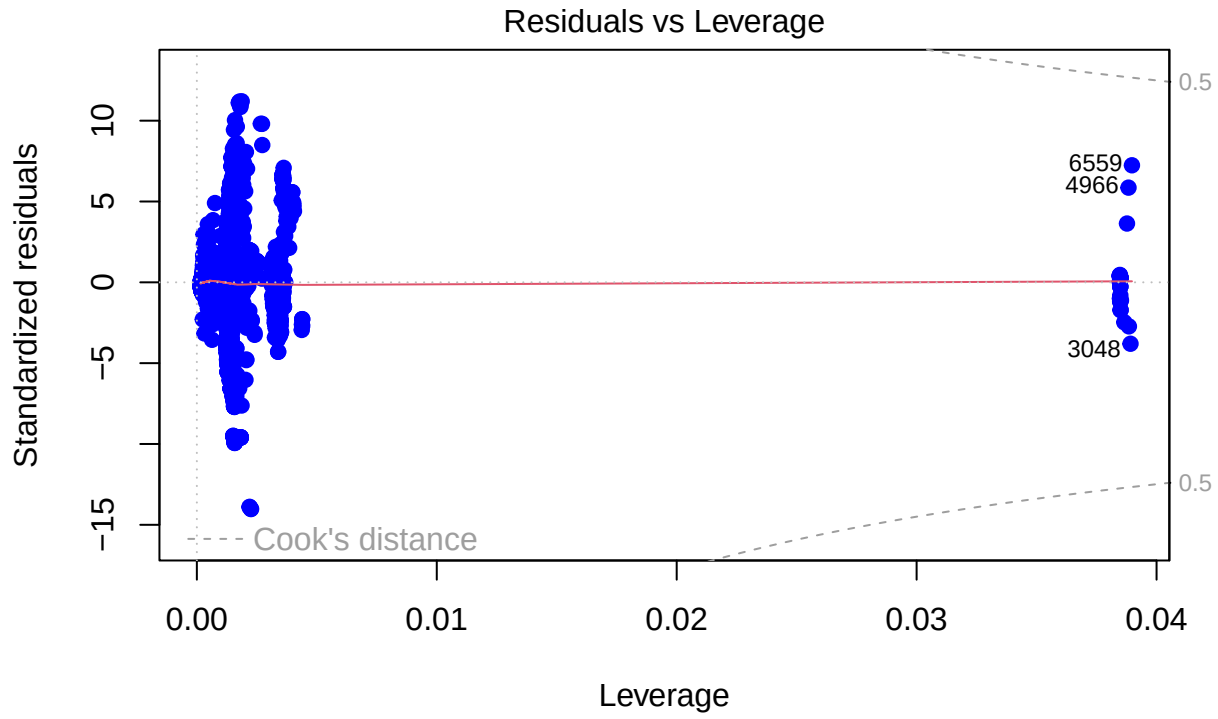
```
##
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS +
##     TRANSMISSION + ENGINE, data = co2_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.248  -1.279  -0.135   1.029  52.924
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  -1.707e+02  1.389e+01  -12.291  < 2e-16 ***
## CONSUMPTION                   2.288e+01  2.157e-02 1061.064  < 2e-16 ***
```

```
## FUELEtanol                    -1.511e+02  3.340e-01 -452.445  < 2e-16 ***
## FUELNaturalGas                -1.033e+02  9.705e-01 -106.430  < 2e-16 ***
## FUELPremium                   -3.515e+01  2.706e-01 -129.916  < 2e-16 ***
## FUELRegular                   -3.463e+01  2.700e-01 -128.237  < 2e-16 ***
## YEAR                           1.022e-01  6.961e-03   14.680  < 2e-16 ***
## GEARS                          6.622e-01  3.699e-02   17.902  < 2e-16 ***
## TRANSMISSIONAutomatic         -6.357e-01  1.811e-01   -3.511 0.000447 ***
## TRANSMISSIONAutomaticSShift   -6.980e-01  1.752e-01   -3.984 6.81e-05 ***
## TRANSMISSIONContinuosVariable  1.552e+00  2.977e-01    5.213 1.87e-07 ***
## TRANSMISSIONManual            -9.862e-01  1.830e-01   -5.388 7.20e-08 ***
## ENGINE                        -7.035e-02  4.317e-02   -1.629 0.103227
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.731 on 20847 degrees of freedom
## Multiple R-squared:  0.9944, Adjusted R-squared:  0.9944
## F-statistic: 3.069e+05 on 12 and 20847 DF,  p-value: < 2.2e-16
```

5.2.4.1 Detecting unequal Variance




Q−Q Residuals
Standardized residuals
10
5
0
−5
−10
−15
−4
−2
0
2
4
Theoretical Quantiles
m(EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION + EN(
Scale−Location
√|Standardized residuals|
3
2
1
0
100
200
300
400
500
600
Fitted values
m(EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION + EN(



Insights

- In the residual plot, it can be seen a pattern out of the 2*REE threshold.

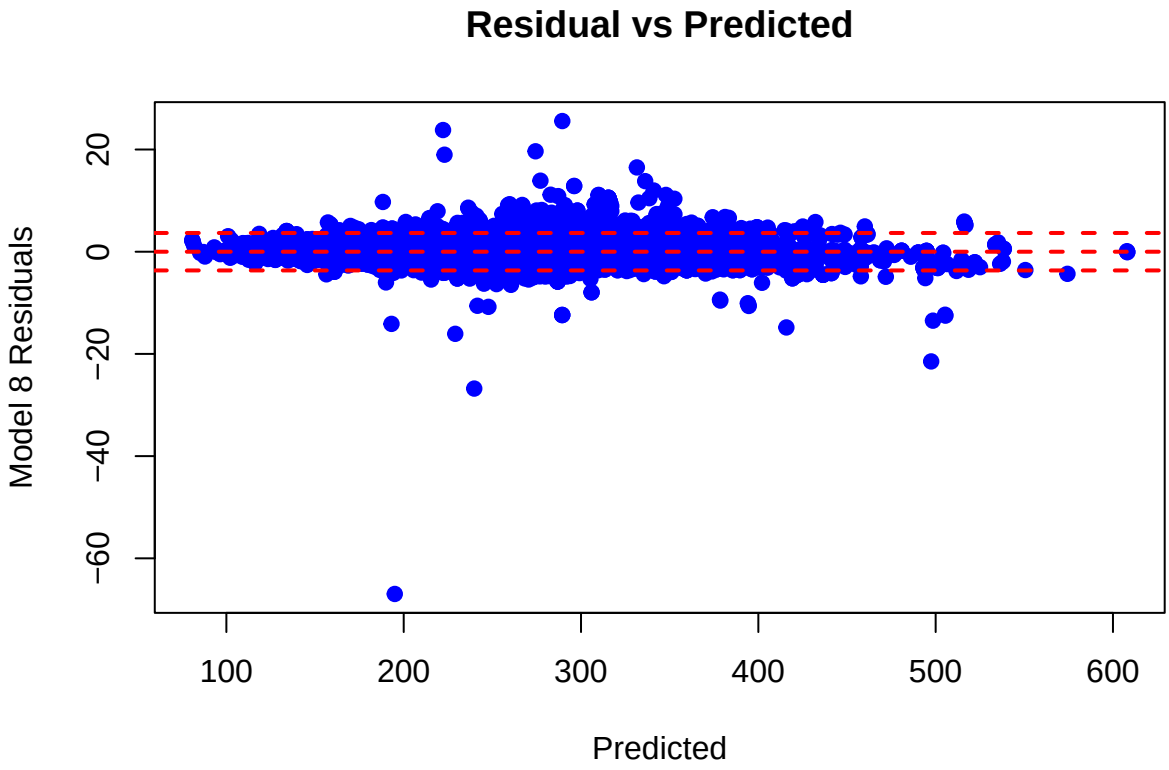
5.2.5 Model 8: Model 7 plus Interaction

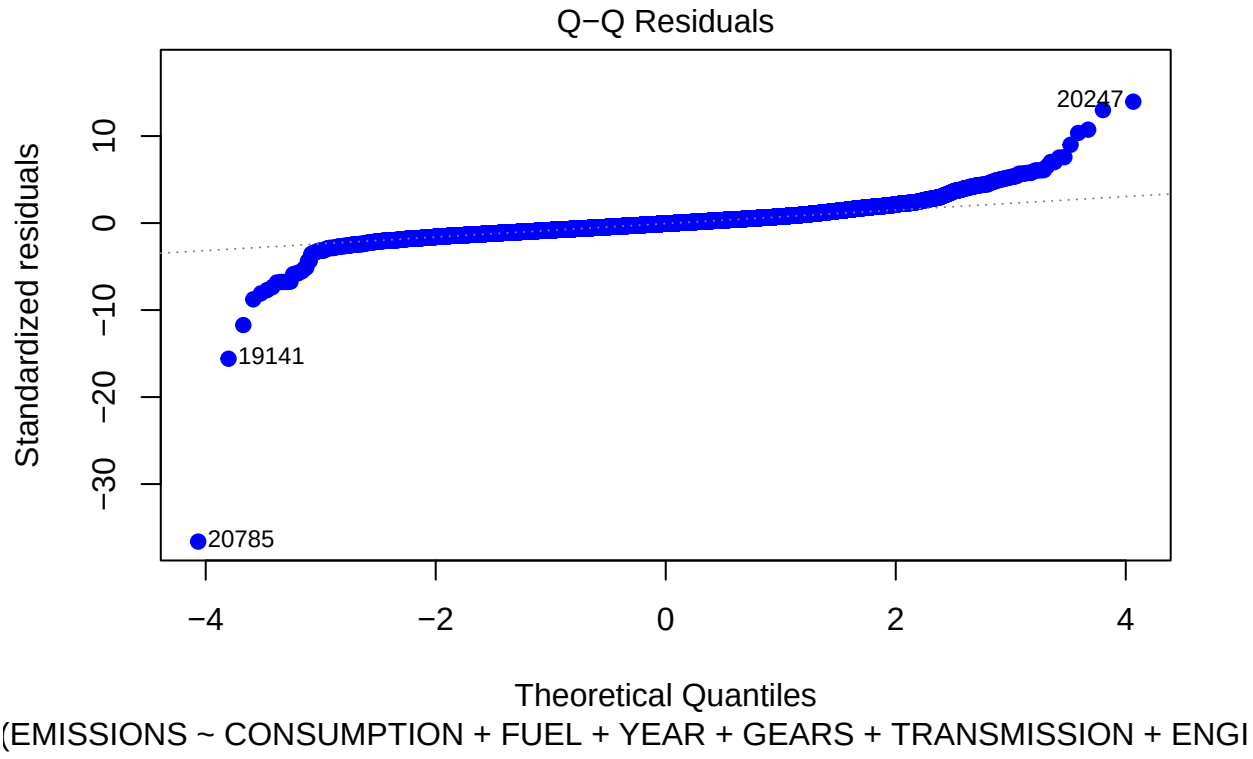
```
model8 = lm(data=co2_train,
            formula = EMISSIONS ~CONSUMPTION+FUEL+YEAR+GEARS+TRANSMISSION+ENGINE+CONSUMPTION*FUEL)
s_model8 = summary(model8)
s_model8
```

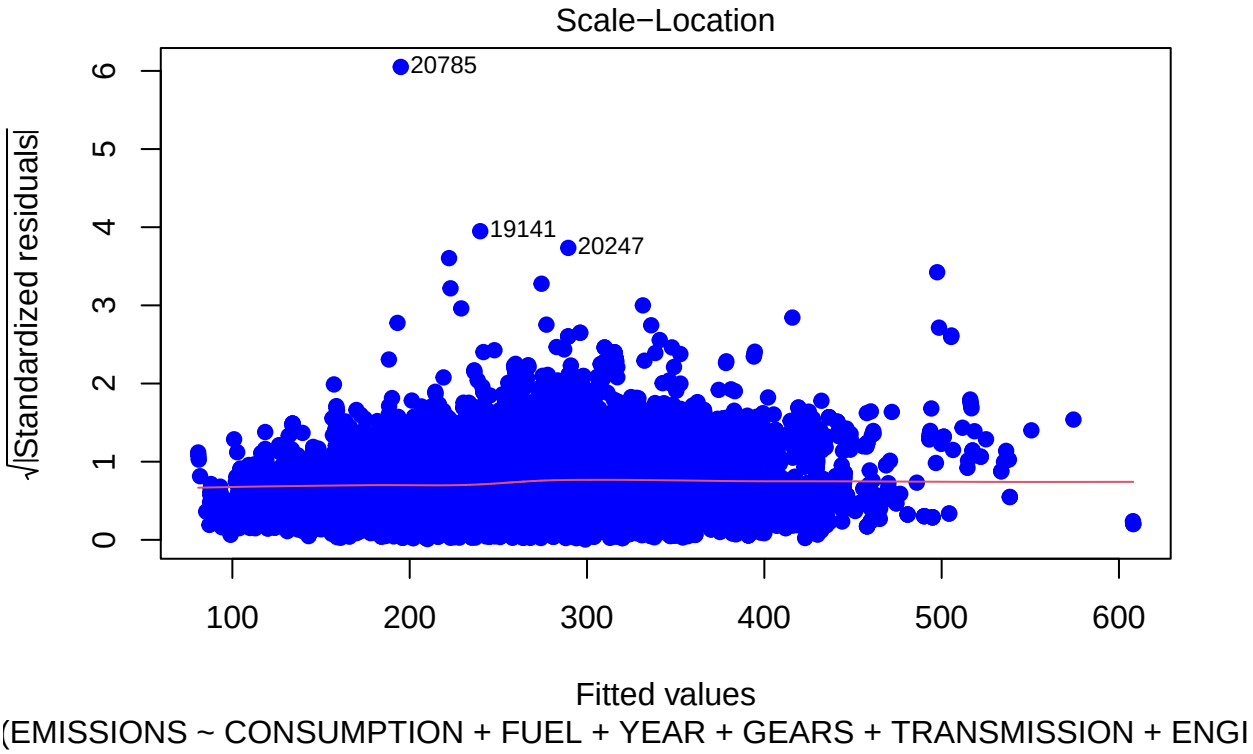
```
##
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS +
##     TRANSMISSION + ENGINE + CONSUMPTION * FUEL, data = co2_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.969  -1.065  -0.110   0.854  25.564
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  -3.639e+02  5.426e+00  -67.078  < 2e-16 ***
## CONSUMPTION                   2.732e+01  3.649e-02  748.607  < 2e-16 ***
## FUELEtanol                    9.815e-01  5.042e-01    1.947  0.05158 .
## FUELNaturalGas               -6.717e+00  2.246e+00   -2.990  0.00279 **
## FUELPremium                   6.546e-01  3.617e-01    1.810  0.07029 .
## FUELRegular                   1.506e+00  3.575e-01    4.212 2.54e-05 ***
## YEAR                          1.791e-01  2.713e-03   65.994  < 2e-16 ***
## GEARS                         3.677e-01  1.437e-02   25.597  < 2e-16 ***
```

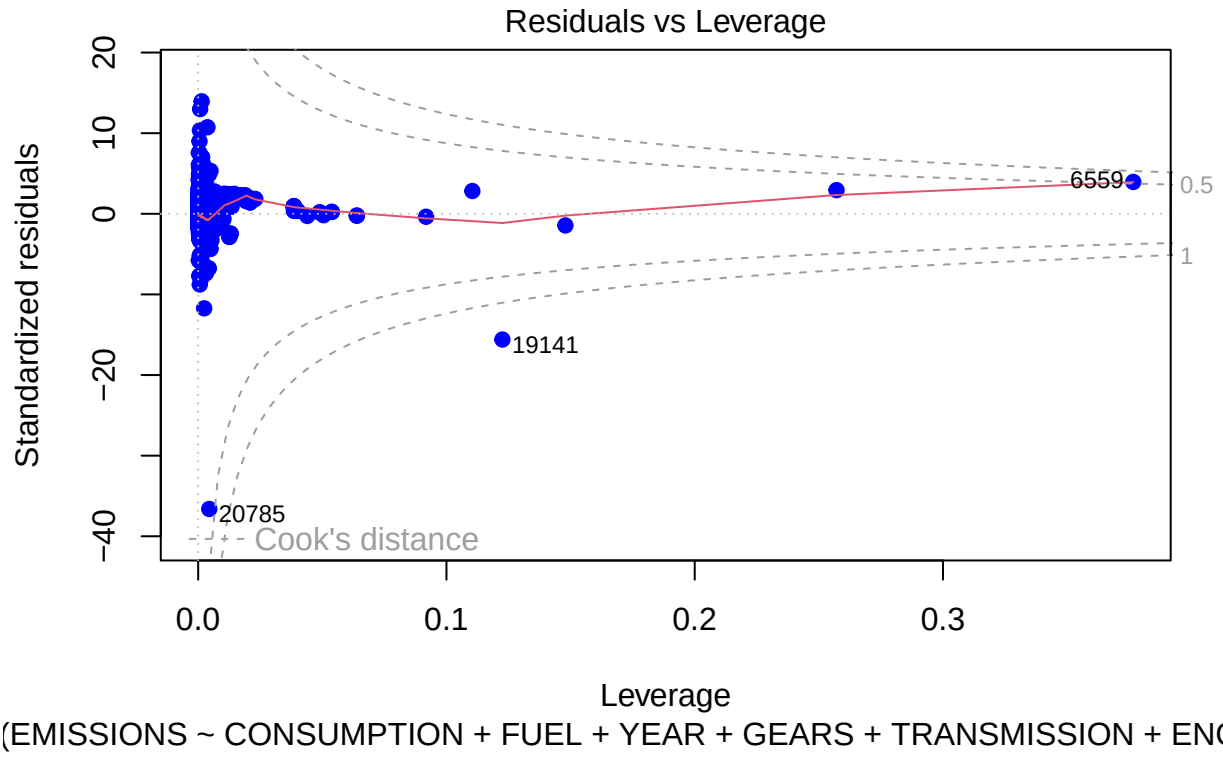
```
## TRANSMISSIONAutomatic          -3.043e-01  7.030e-02   -4.329 1.51e-05 ***
## TRANSMISSIONAutomaticSShift    -5.601e-01  6.792e-02   -8.247  < 2e-16 ***
## TRANSMISSIONContinuosVariable  1.329e+00  1.155e-01   11.509  < 2e-16 ***
## TRANSMISSIONManual             -4.365e-01  7.097e-02   -6.151 7.82e-10 ***
## ENGINE                         -5.027e-01  1.682e-02  -29.890  < 2e-16 ***
## CONSUMPTION:FUELEtanol         -1.101e+01  4.166e-02 -264.197  < 2e-16 ***
## CONSUMPTION:FUELNaturalGas     -7.663e+00  1.354e-01  -56.585  < 2e-16 ***
## CONSUMPTION:FUELPremium        -3.917e+00  3.656e-02 -107.147  < 2e-16 ***
## CONSUMPTION:FUELRegular        -3.953e+00  3.626e-02 -109.029  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.833 on 20843 degrees of freedom
## Multiple R-squared:  0.9992, Adjusted R-squared:  0.9992
## F-statistic: 1.54e+06 on 16 and 20843 DF,  p-value: < 2.2e-16
```

5.2.5.1 Detecting unequal Variance




Q−Q Residuals
Standardized residuals
10
0
−10
−20
−30
20247
19141
20785
−4
−2
0
2
4
Theoretical Quantiles
(EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION + ENGI


Scale−Location
√|Standardized residuals|
6
5
4
3
2
1
0
20785
19141
20247
100
200
300
400
500
600
Fitted values
(EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS + TRANSMISSION + ENGI



Insights

- The term with the interaction eliminated the pattern in the residual plot.
- The model has a high r-squared, a small standard error, and all the coefficients are significant.

6 Modeling based on Descriptive Analysis

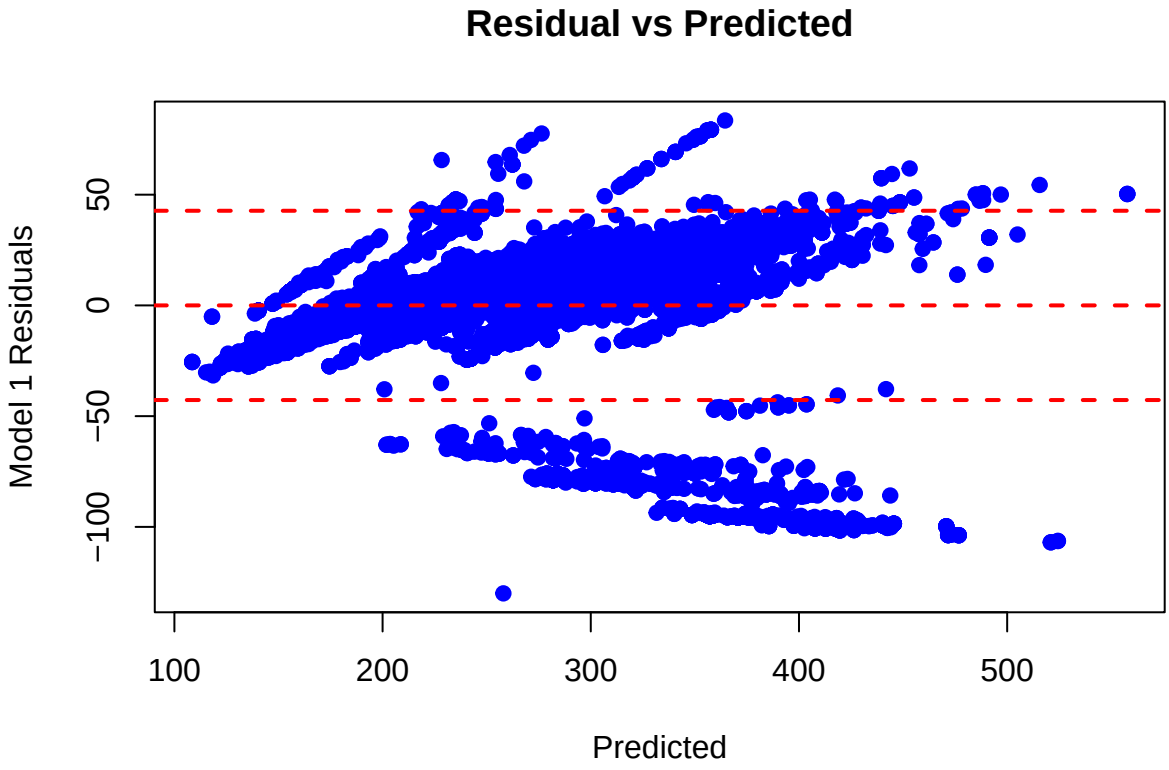
6.1 Model 1: Numerical Variables

```
model1 = lm(data = co2_train,
            formula = EMISSIONS~CONSUMPTION+CYLINDERS+ENGINE)
s_model1 = summary(model1)
s_model1
```

```
##
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + CYLINDERS + ENGINE, data = co2_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -130.015   -4.710    2.010    9.593   83.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.00429    0.66105   51.44   <2e-16 ***
## CONSUMPTION 17.00236    0.08187  207.68   <2e-16 ***
## CYLINDERS    3.33742    0.19997   16.69   <2e-16 ***
## ENGINE       3.31655    0.28966   11.45   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 21.36 on 20856 degrees of freedom
## Multiple R-squared:  0.8851, Adjusted R-squared:  0.8851
## F-statistic: 5.357e+04 on 3 and 20856 DF,  p-value: < 2.2e-16
```

6.1.1 Detecting unequal Variance



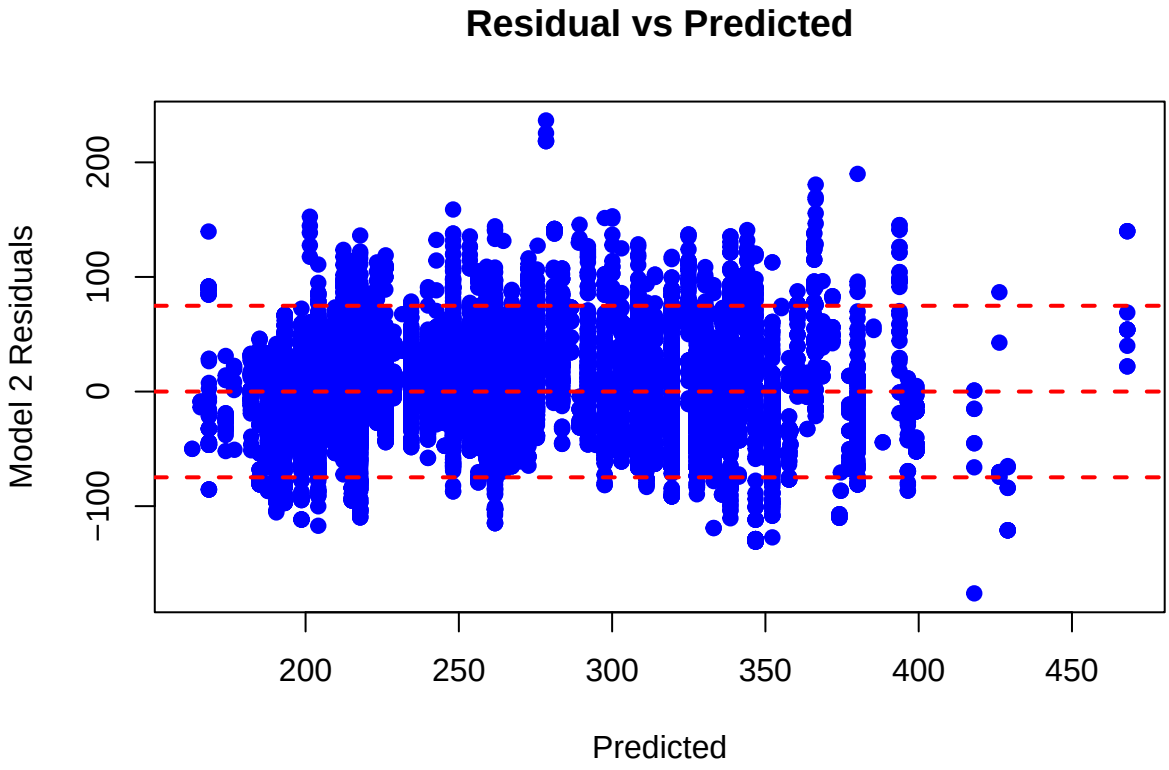
6.2 Model 2: Numerical Variables without Consumption

```r
model2 = lm(data = co2_train,
            formula = EMISSIONS~CYLINDERS+ENGINE)
s_model2 = summary(model2)
s_model2
```

```
##
## Call:
## lm(formula = EMISSIONS ~ CYLINDERS + ENGINE, data = co2_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -176.120  -24.208   -2.656   22.328  236.559
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 116.1766     0.9275  125.25   <2e-16 ***
## CYLINDERS     8.3211     0.3477   23.93   <2e-16 ***
## ENGINE       27.3415     0.4651   58.78   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 37.42 on 20857 degrees of freedom
## Multiple R-squared:  0.6476, Adjusted R-squared:  0.6475
## F-statistic: 1.916e+04 on 2 and 20857 DF,  p-value: < 2.2e-16
```

6.2.1 Detecting unequal Variance



6.3 Model 3: Numerical and Categorical

```
model3 = lm(data = co2_train,
            formula=EMISSIONS~CONSUMPTION+CYLINDERS+ENGINE+FUEL+TRANSMISSION)
s_model3 = summary(model3)
s_model3
```

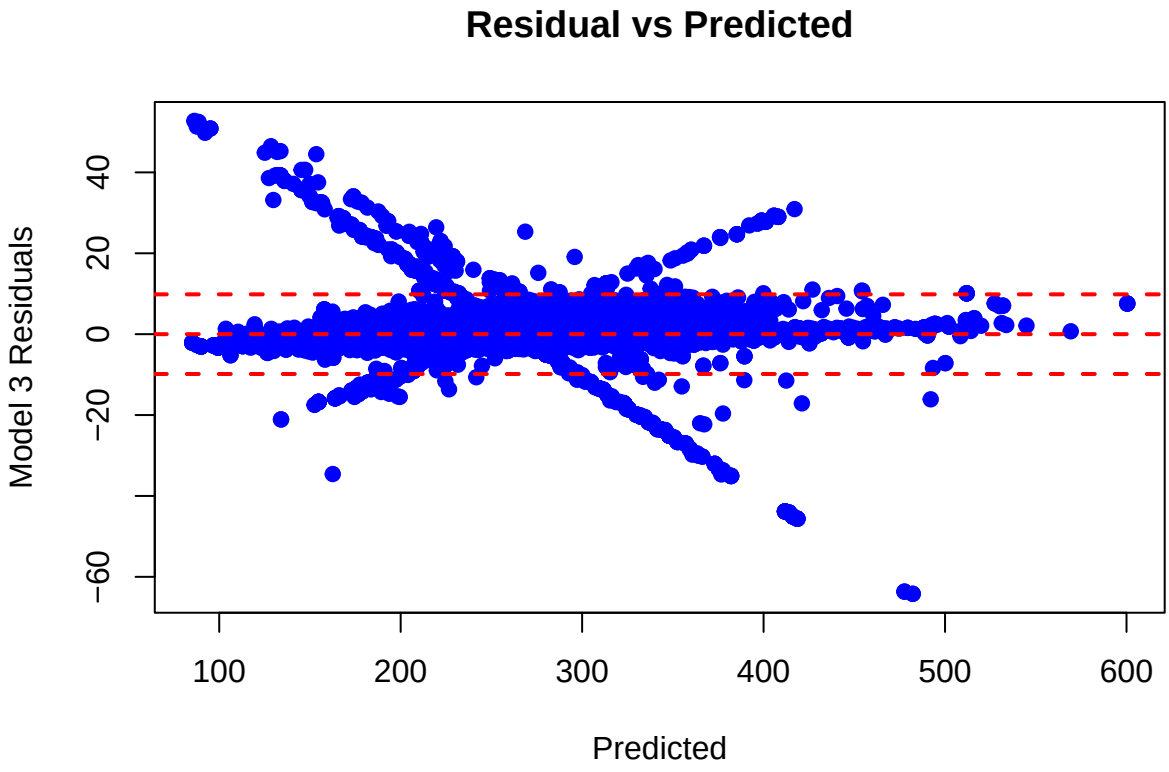
```
##
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + CYLINDERS + ENGINE + FUEL +
##     TRANSMISSION, data = co2_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.196  -1.433  -0.630   0.830  52.724
##
## Coefficients:
##                               Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                   41.25178    0.35260  116.992  < 2e-16 ***
## CONSUMPTION                   22.70103    0.02190 1036.456  < 2e-16 ***
## CYLINDERS                      0.01668    0.04938    0.338  0.73560
```

```
## ENGINE                           0.21394    0.06898    3.101  0.00193 **
## FUELEtanol                    -150.43381    0.34118 -440.924  < 2e-16 ***
## FUELNaturalGas                -104.65314    1.00839 -103.783  < 2e-16 ***
## FUELPremium                    -35.28884    0.28136 -125.422  < 2e-16 ***
## FUELRegular                    -35.40093    0.27815 -127.275  < 2e-16 ***
## TRANSMISSIONAutomatic           -2.53600    0.18159  -13.966  < 2e-16 ***
## TRANSMISSIONAutomaticSShift     -1.14143    0.18117   -6.300 3.03e-10 ***
## TRANSMISSIONContinuosVariable   -2.52036    0.23978  -10.511  < 2e-16 ***
## TRANSMISSIONManual              -2.73779    0.18394  -14.884  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.922 on 20848 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9939
## F-statistic: 3.091e+05 on 11 and 20848 DF,  p-value: < 2.2e-16
```

Insights

- Cylinders is not significant, because of the high correlation with engine.

6.3.1 Detecting unequal Variance



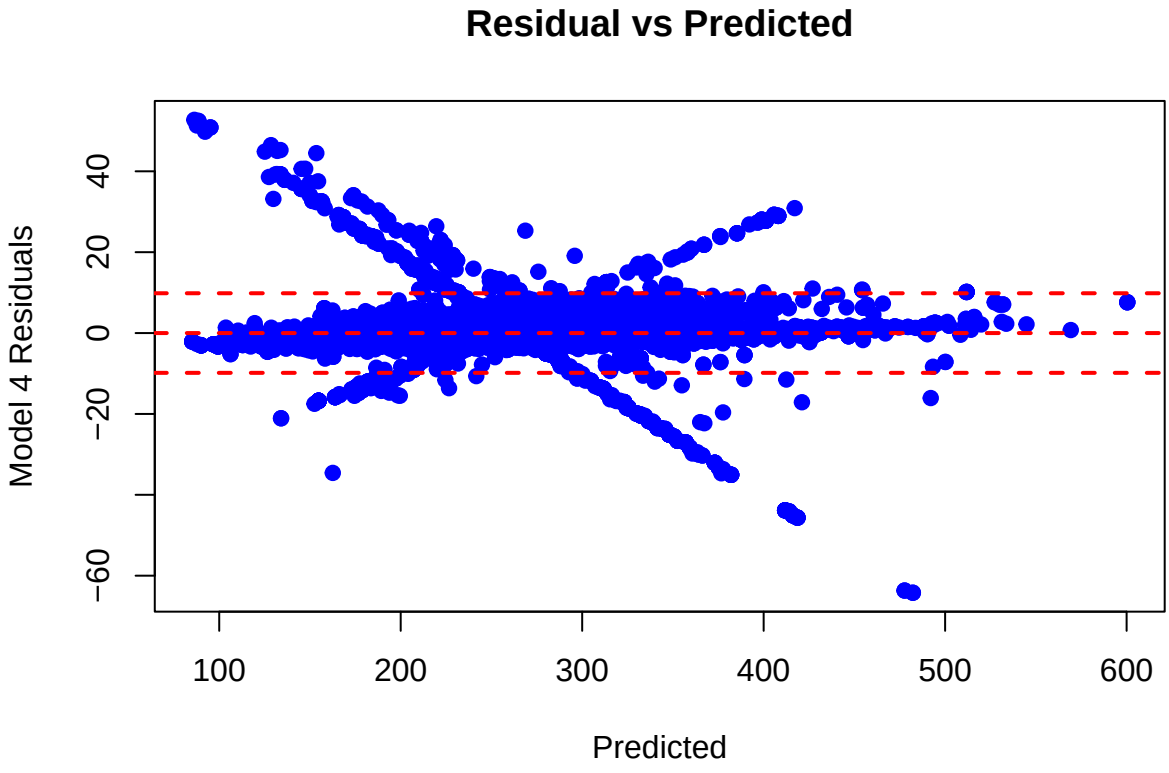
6.4 Model 4: Model3 without Cylinders

```
model4 = lm(data = co2_train,
            formula=EMISSIONS~CONSUMPTION+ENGINE+FUEL+TRANSMISSION)
s_model4 = summary(model4)
s_model4
```

```
##
```

```
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + ENGINE + FUEL + TRANSMISSION,
##     data = co2_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.218  -1.430  -0.630   0.825  52.728
##
## Coefficients:
##                                 Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                     41.27969    0.34278  120.427  < 2e-16 ***
## CONSUMPTION                     22.70222    0.02161 1050.401  < 2e-16 ***
## ENGINE                           0.23186    0.04408    5.260 1.46e-07 ***
## FUELEtanol                    -150.44115    0.34048 -441.852  < 2e-16 ***
## FUELNaturalGas                -104.66562    1.00769 -103.867  < 2e-16 ***
## FUELPremium                    -35.28394    0.28098 -125.575  < 2e-16 ***
## FUELRegular                    -35.40427    0.27796 -127.370  < 2e-16 ***
## TRANSMISSIONAutomatic           -2.53977    0.18124  -14.013  < 2e-16 ***
## TRANSMISSIONAutomaticSShift     -1.14454    0.18093   -6.326 2.57e-10 ***
## TRANSMISSIONContinuosVariable   -2.52406    0.23952  -10.538  < 2e-16 ***
## TRANSMISSIONManual              -2.74310    0.18326  -14.968  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.922 on 20849 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9939
## F-statistic: 3.4e+05 on 10 and 20849 DF,  p-value: < 2.2e-16
```

6.4.1 Detecting unequal Variance



6.5 Model 5: Model4+Interaction

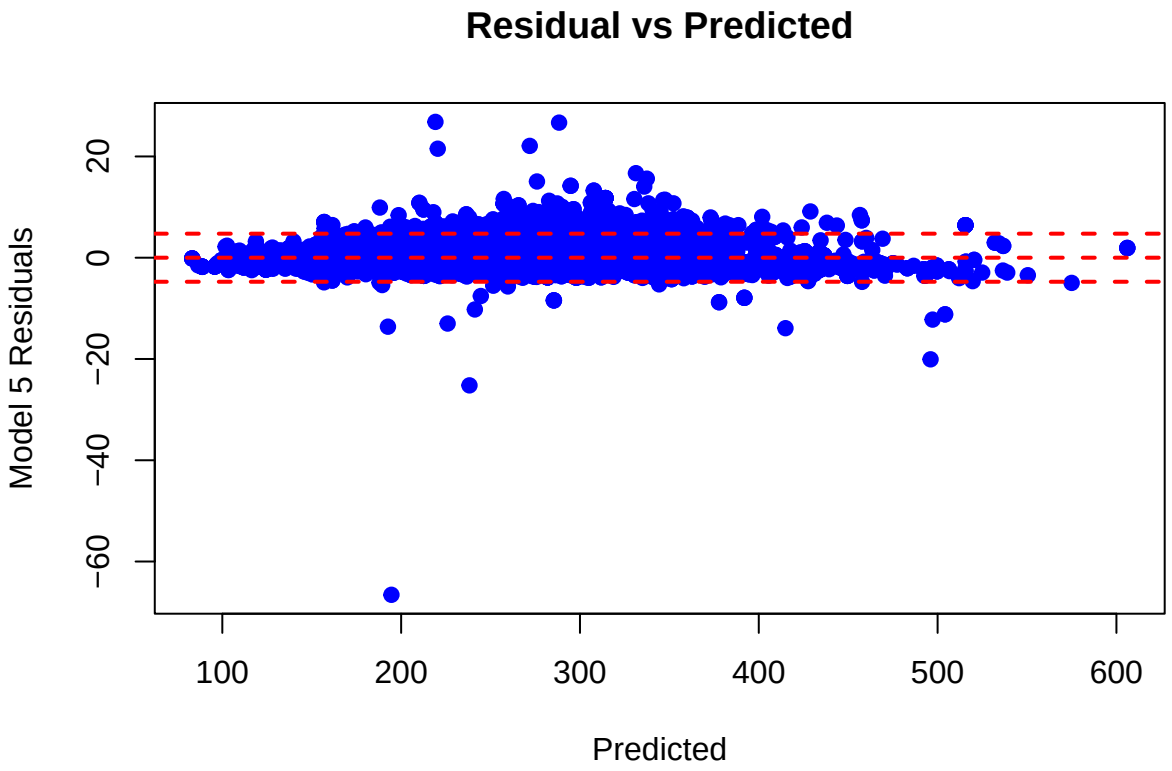
```
model5 = lm(data = co2_train,
            formula=EMISSIONS~CONSUMPTION+ENGINE+FUEL+TRANSMISSION+CONSUMPTION*FUEL)
s_model5 = summary(model5)
s_model5
```

```
##
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + ENGINE + FUEL + TRANSMISSION +
##     CONSUMPTION * FUEL, data = co2_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.570  -1.269  -0.704   0.460  26.823
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                     1.39748    0.46032    3.036  0.00240 **
## CONSUMPTION                    27.07171    0.04714  574.231  < 2e-16 ***
## ENGINE                         -0.13952    0.02138   -6.526 6.93e-11 ***
## FUELEtanol                      0.41456    0.65268    0.635  0.52533
## FUELNaturalGas                 -8.97395    2.90822   -3.086  0.00203 **
## FUELPremium                     0.11155    0.46815    0.238  0.81166
## FUELRegular                     1.35406    0.46234    2.929  0.00341 **
## TRANSMISSIONAutomatic          -2.51424    0.08769  -28.671  < 2e-16 ***
## TRANSMISSIONAutomaticSShift    -1.21899    0.08730  -13.964  < 2e-16 ***
## TRANSMISSIONContinuosVariable  -1.46664    0.11591  -12.654  < 2e-16 ***
## TRANSMISSIONManual             -2.59580    0.08845  -29.349  < 2e-16 ***
## CONSUMPTION:FUELEtanol        -10.88368    0.05387 -202.022  < 2e-16 ***
## CONSUMPTION:FUELNaturalGas     -7.58262    0.17534  -43.246  < 2e-16 ***
## CONSUMPTION:FUELPremium        -3.86530    0.04733  -81.668  < 2e-16 ***
## CONSUMPTION:FUELRegular        -3.98229    0.04694  -84.833  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.374 on 20845 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.049e+06 on 14 and 20845 DF,  p-value: < 2.2e-16
```

INSIGHTS

- Engine turned negative

6.5.1 Detecting unequal Variance



6.6 Model 6: Model5-Engine

```
model6 = lm(data = co2_train,
            formula=EMISSIONS~CONSUMPTION+FUEL+TRANSMISSION+CONSUMPTION*FUEL)
s_model6 = summary(model6)
s_model6
```

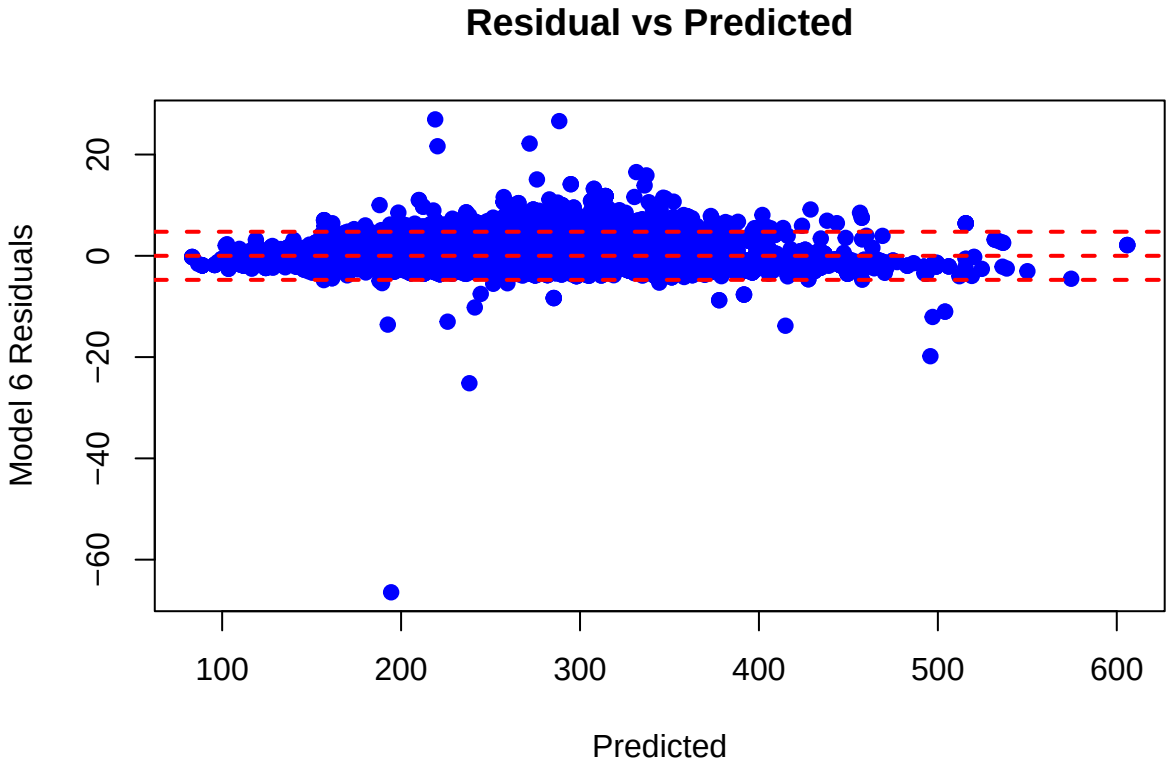
```
##
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + FUEL + TRANSMISSION +
##     CONSUMPTION * FUEL, data = co2_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.442  -1.261  -0.701   0.468  26.938
##
## Coefficients:
##                                Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                     1.55471    0.46014    3.379  0.00073 ***
## CONSUMPTION                    27.01290    0.04632  583.168  < 2e-16 ***
## FUELEtanol                      0.39252    0.65332    0.601  0.54797
## FUELNaturalGas                 -9.07724    2.91108   -3.118  0.00182 **
## FUELPremium                     0.15054    0.46858    0.321  0.74800
## FUELRegular                     1.33542    0.46279    2.886  0.00391 **
## TRANSMISSIONAutomatic          -2.55002    0.08761  -29.107  < 2e-16 ***
## TRANSMISSIONAutomaticSShift    -1.25108    0.08724  -14.340  < 2e-16 ***
## TRANSMISSIONContinuosVariable  -1.54475    0.11540  -13.386  < 2e-16 ***
```

```
## TRANSMISSIONManual             -2.59199    0.08853  -29.277  < 2e-16 ***
## CONSUMPTION:FUELEtanol        -10.86745    0.05387 -201.734  < 2e-16 ***
## CONSUMPTION:FUELNaturalGas     -7.56915    0.17550  -43.130  < 2e-16 ***
## CONSUMPTION:FUELPremium        -3.86477    0.04738  -81.576  < 2e-16 ***
## CONSUMPTION:FUELRegular        -3.97446    0.04697  -84.609  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.376 on 20846 degrees of freedom
## Multiple R-squared:  0.9986, Adjusted R-squared:  0.9986
## F-statistic: 1.128e+06 on 13 and 20846 DF,  p-value: < 2.2e-16
```

Insights

- Some fuel categories became not significant by their own (without interaction).

6.6.1 Detecting unequal Variance



6.7 Model 9: Simple Model

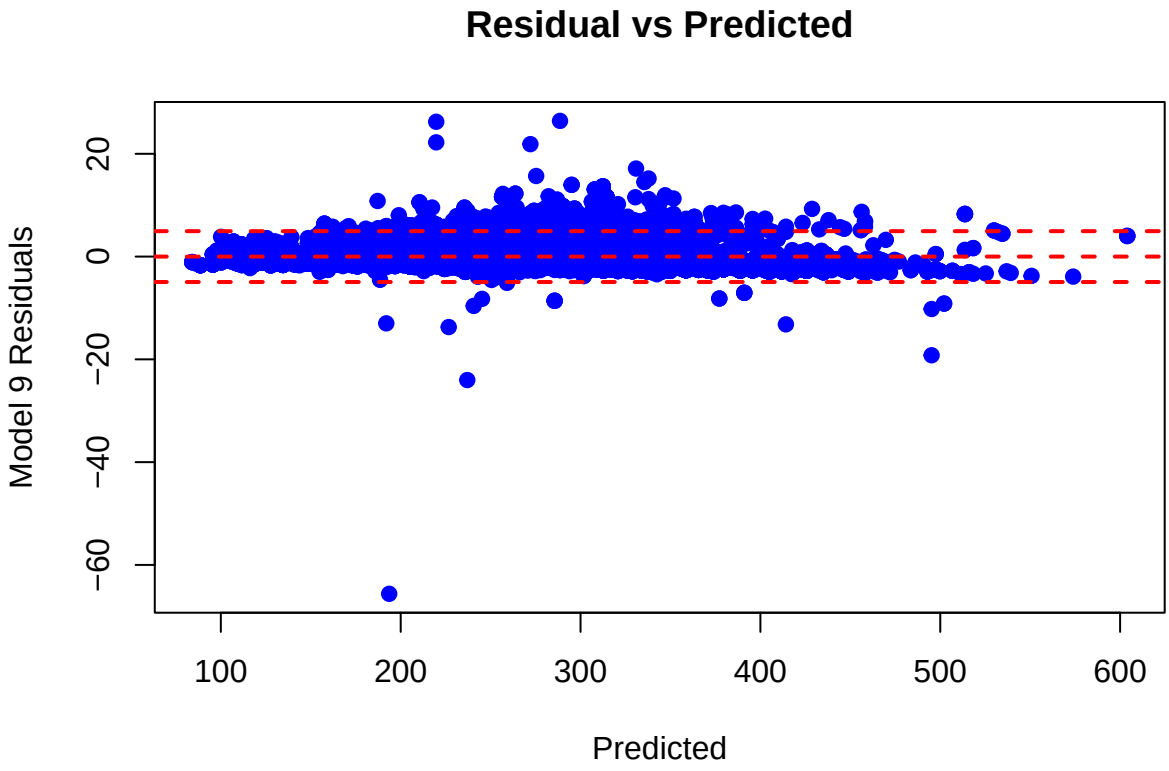
```
model9 = lm(data = co2_train,
            formula=EMISSIONS~CONSUMPTION+FUEL+CONSUMPTION*FUEL)
s_model9 = summary(model9)
s_model9
```

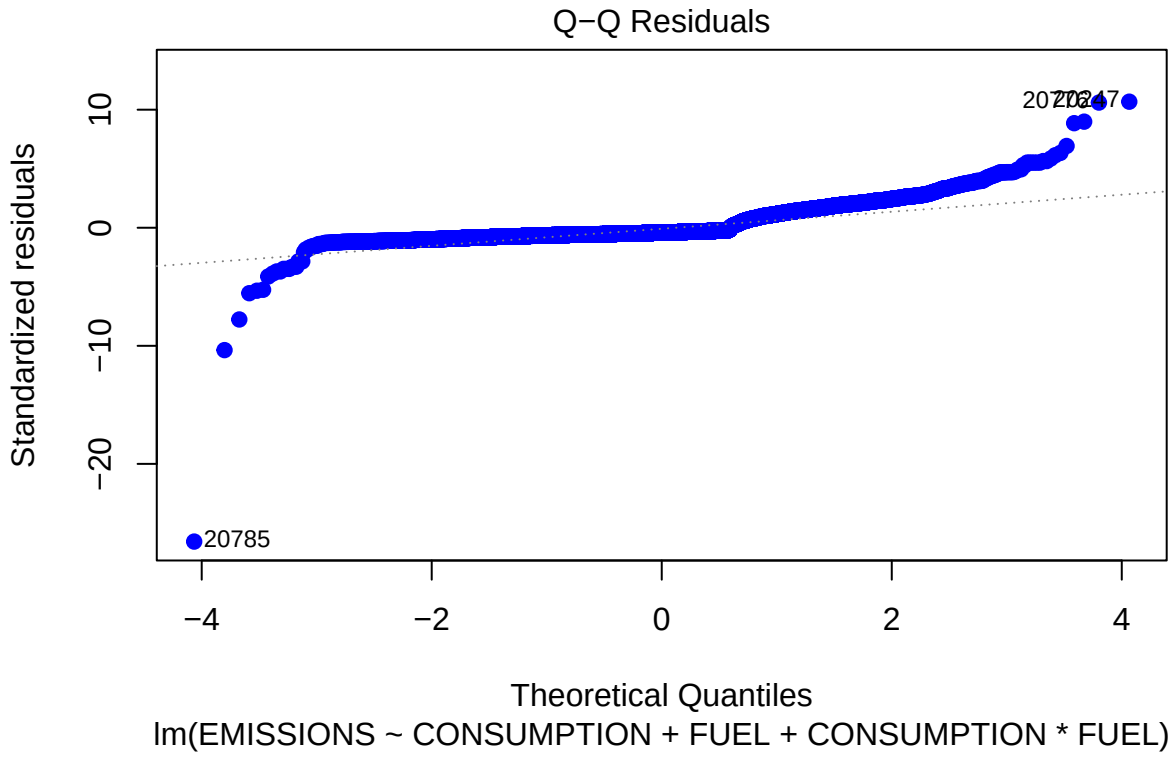
```
##
## Call:
## lm(formula = EMISSIONS ~ CONSUMPTION + FUEL + CONSUMPTION * FUEL,
##     data = co2_train)
##
```

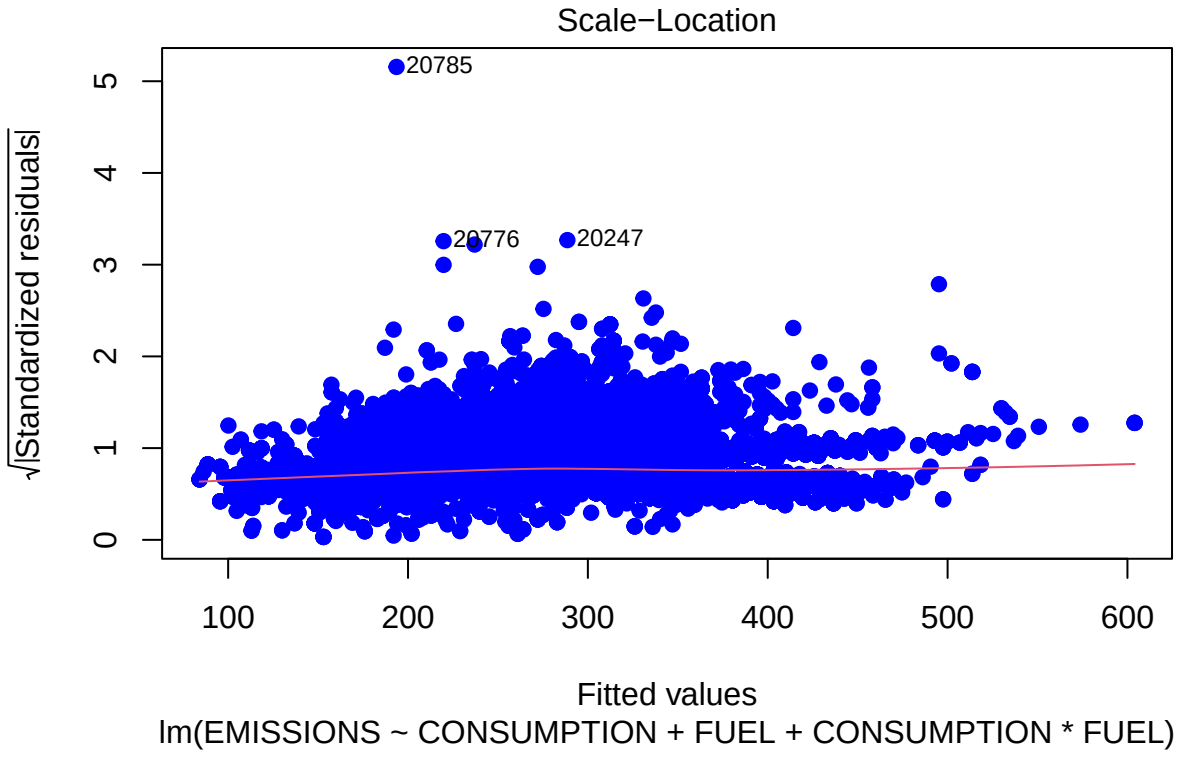
```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -65.617  -1.418  -1.025   0.987  26.400
##
## Coefficients:
##                              Estimate Std. Error  t value Pr(>|t|)
## (Intercept)                  -0.23092    0.47187   -0.489 0.624586
## CONSUMPTION                  26.96563    0.04815  559.982  < 2e-16 ***
## FUELEtanol                    0.66427    0.68014    0.977 0.328742
## FUELNaturalGas               -9.34275    3.03095   -3.082 0.002056 **
## FUELPremium                   0.11727    0.48782    0.240 0.810020
## FUELRegular                   1.59207    0.48067    3.312 0.000927 ***
## CONSUMPTION:FUELEtanol      -10.86696    0.05608 -193.774  < 2e-16 ***
## CONSUMPTION:FUELNaturalGas  -7.54866    0.18272  -41.312  < 2e-16 ***
## CONSUMPTION:FUELPremium     -3.82005    0.04930  -77.488  < 2e-16 ***
## CONSUMPTION:FUELRegular     -3.98968    0.04885  -81.679  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.474 on 20850 degrees of freedom
## Multiple R-squared:  0.9985, Adjusted R-squared:  0.9985
## F-statistic: 1.502e+06 on 9 and 20850 DF,  p-value: < 2.2e-16
```

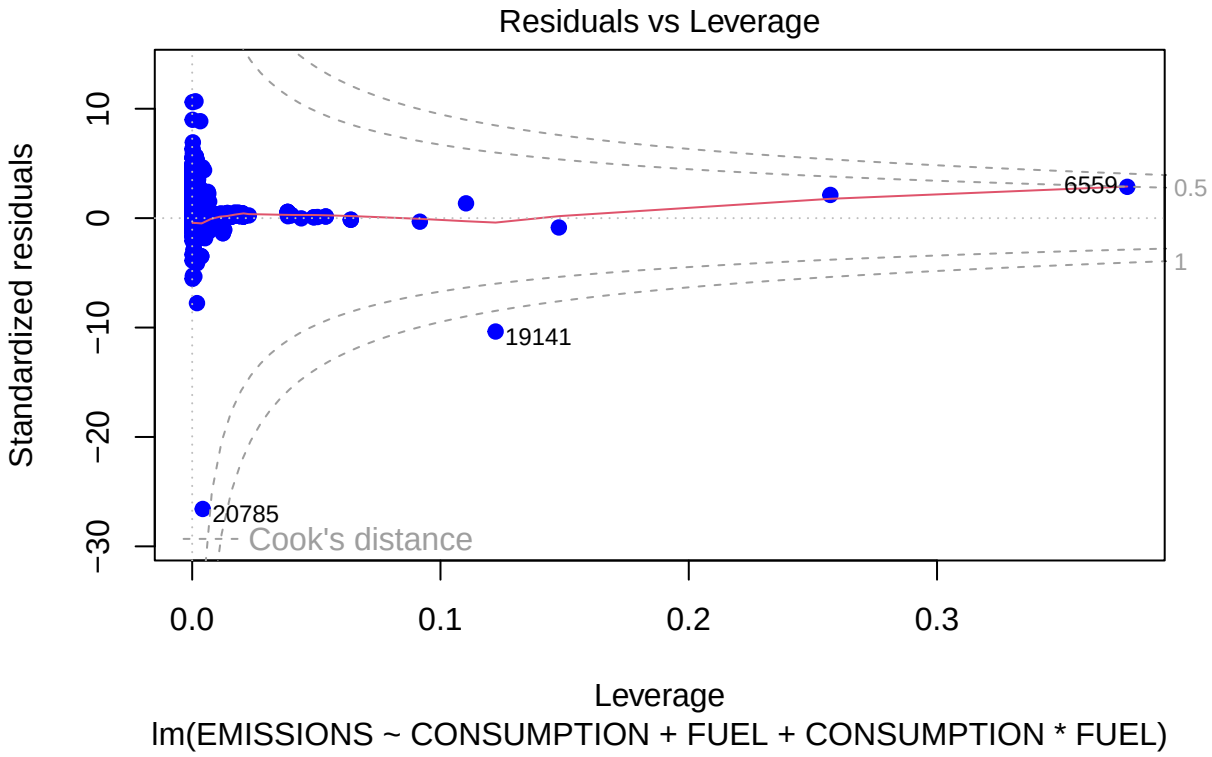
```
## The following objects are masked from co2_train (pos = 5):
##
##     CONSUMPTION, CYLINDERS, EMISSIONS, ENGINE, FUEL, GEARS,
##     TRANSMISSION, YEAR
```

6.7.1 Detecting unequal Variance




Q−Q Residuals
Standardized residuals
10
0
−10
−20
20785
−4
−2
0
2
4
Theoretical Quantiles
lm(EMISSIONS ~ CONSUMPTION + FUEL + CONSUMPTION * FUEL)


Scale−Location
√|Standardized residuals|
20785
20776
20247
0
1
2
3
4
5
100
200
300
400
500
600
Fitted values
lm(EMISSIONS ~ CONSUMPTION + FUEL + CONSUMPTION * FUEL)

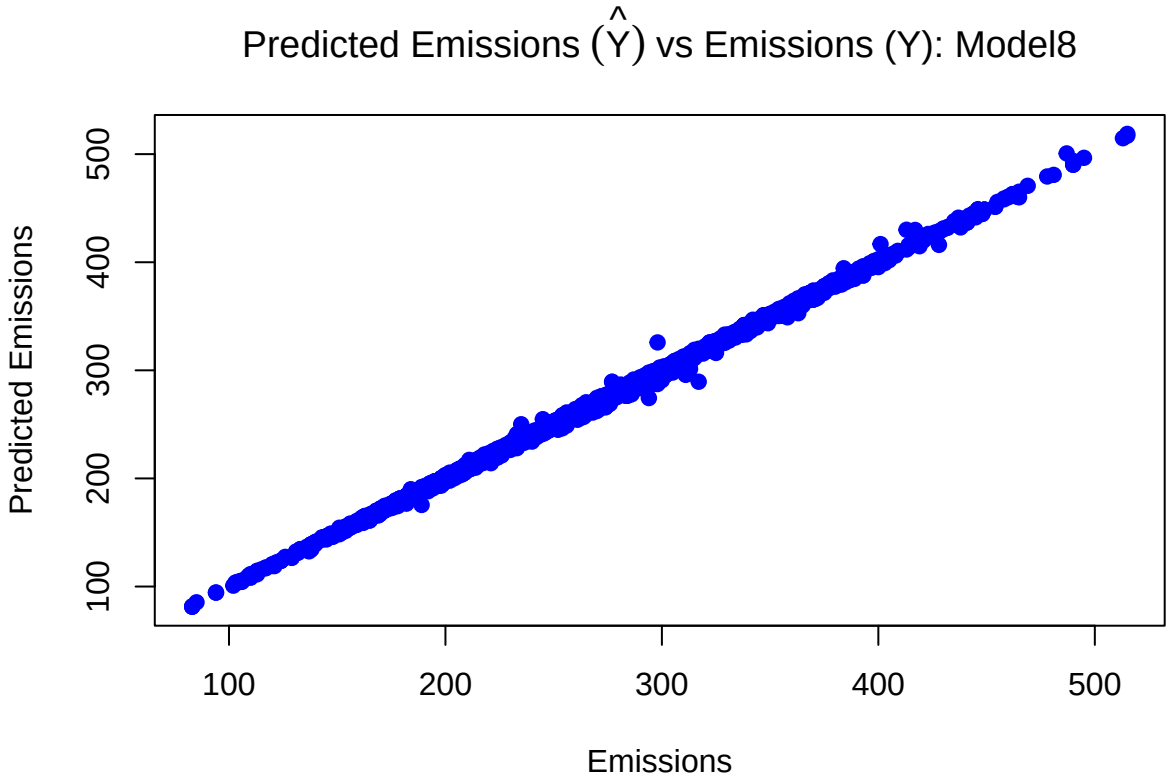


Conclusion

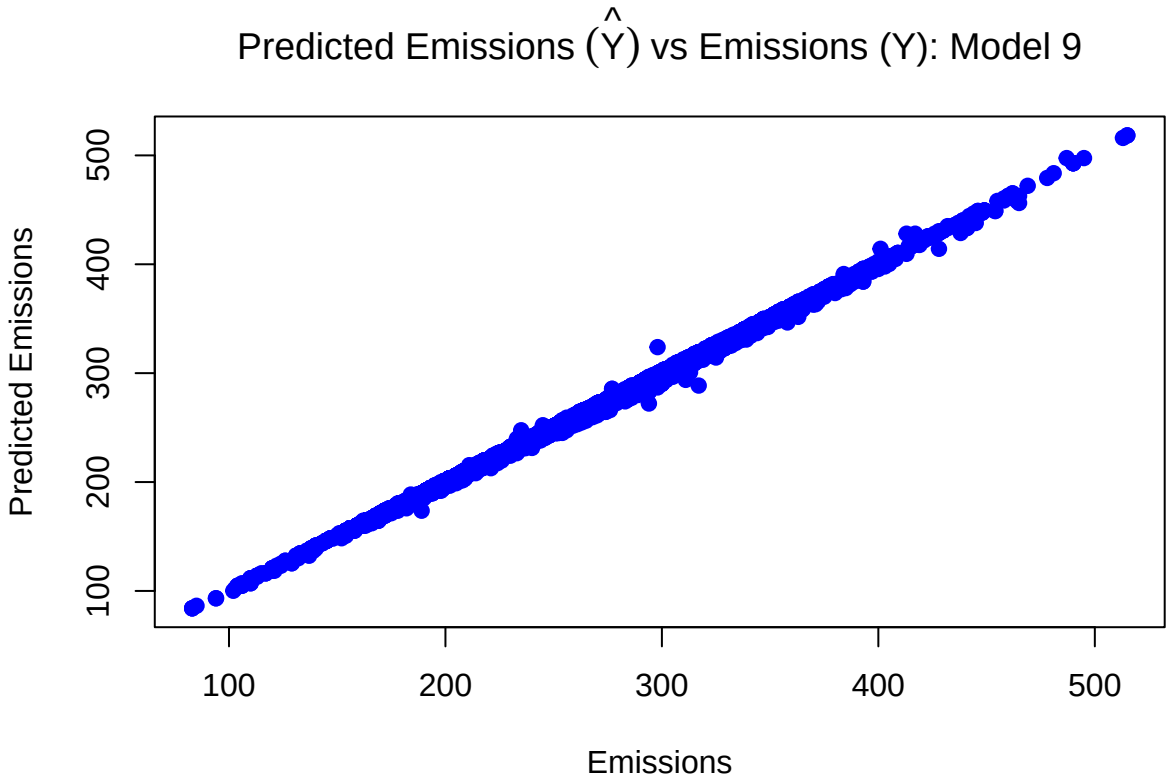
- It was decided to use the model 9 as the chosen model from the descriptive analysis to prioritize the small amount of predictors, and also, model 9 meets of the assumption and has a high adjusted r-square (0.99).

7 Testing

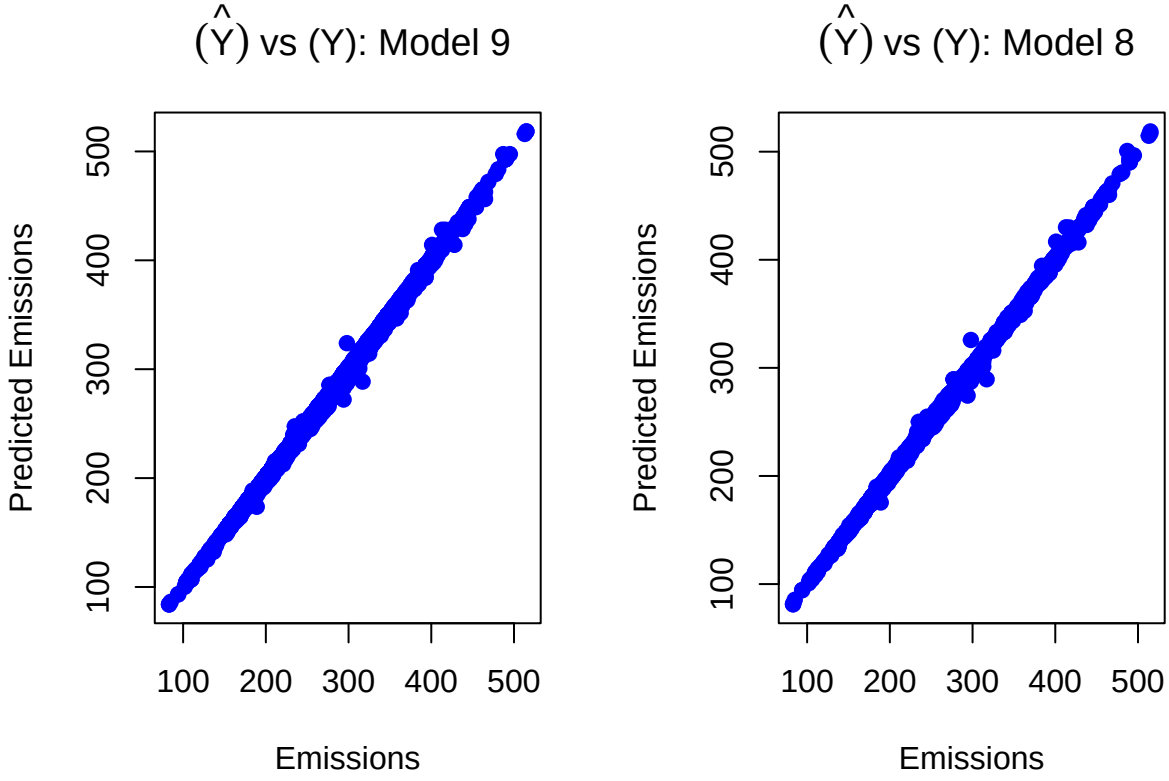
7.1 Model 8



7.2 Model 9



7.3 Comparisson between 8 and 9



8 Conclusions

- The model obtained from the variable screening method had a pattern in the residual plot, which was eliminated with the interaction between consumption and type of fuel (Model8). This relationship was discovered in the descriptive analysis, and when applying to the suggested model improved the r-squared from 0.994 to 0.9992.
- The chosen model based on the information gathered in the descriptive analysis (Model9) only has three coefficients, consumption, type of fuel and its interactions, but still has an adjusted r squared of 0.9985 which is highly relevant.
- As seen in the comparison of the results, when using the testing data, Model 8 and model 9 have similar results which is reasonable as Model 8 has an RMSE of 1.83 and Model 9 has an RMSE of 2.47.
- Therefore, the team select to use Model 9 to predict car emissions for its high accuracy, but above all becuase of is simplicity.