Model Summary

We are interested to see how CO2 Emission responds to model year, make, type, class, engine size, cylinders, transmission, fuel type and the difference in city/hwy fuel consumption, so we will create a dataset for the model.

Part 1 – Merging Data and Cleaning

1. Combined CSV file from 2018 to 2022 in R data frame
2. Rename columns
3. Remove columns that are not relevant: comb lkm, comb mpg, EO2 rating, smog rating.
4. Extract the below info from the model name, then remove the models' name

4WD/4X4 = Four-wheel drive
AWD = All-wheel drive
FFV = Flexible-fuel vehicle
SWB = Short wheelbase
LWB = Long wheelbase
EWB = Extended wheelbase

5. Remove 470 duplicate rows
6. Correct data type: characters to numeric

Output

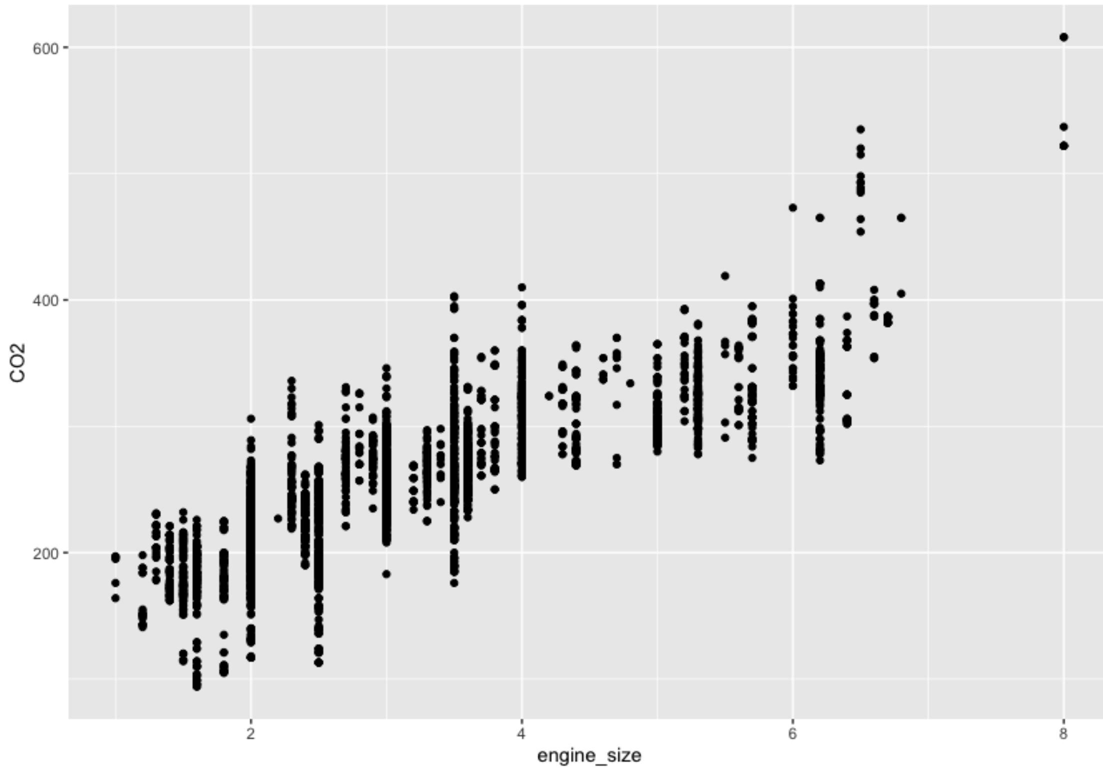
```
> head(df_model)
  year  make      class engine_size cylinders transmission fuel_type fuel_city fuel_hwy CO2 model_type
3 2022 Acura    Compact         2.4         4          AM8         Z       9.9      7.0 200      Other
4 2022 Acura SUV: Small         3.5         6         AS10         Z      12.6      9.4 263        AWD
5 2022 Acura SUV: Small         2.0         4         AS10         Z      11.0      8.6 232        AWD
6 2022 Acura SUV: Small         2.0         4         AS10         Z      11.3      9.1 242        AWD
7 2022 Acura    Compact         2.0         4         AS10         Z      11.2      8.0 230        AWD
8 2022 Acura    Compact         2.0         4         AS10         Z      11.3      8.1 231        AWD
```

Part 2 – EDA and Data Visualization

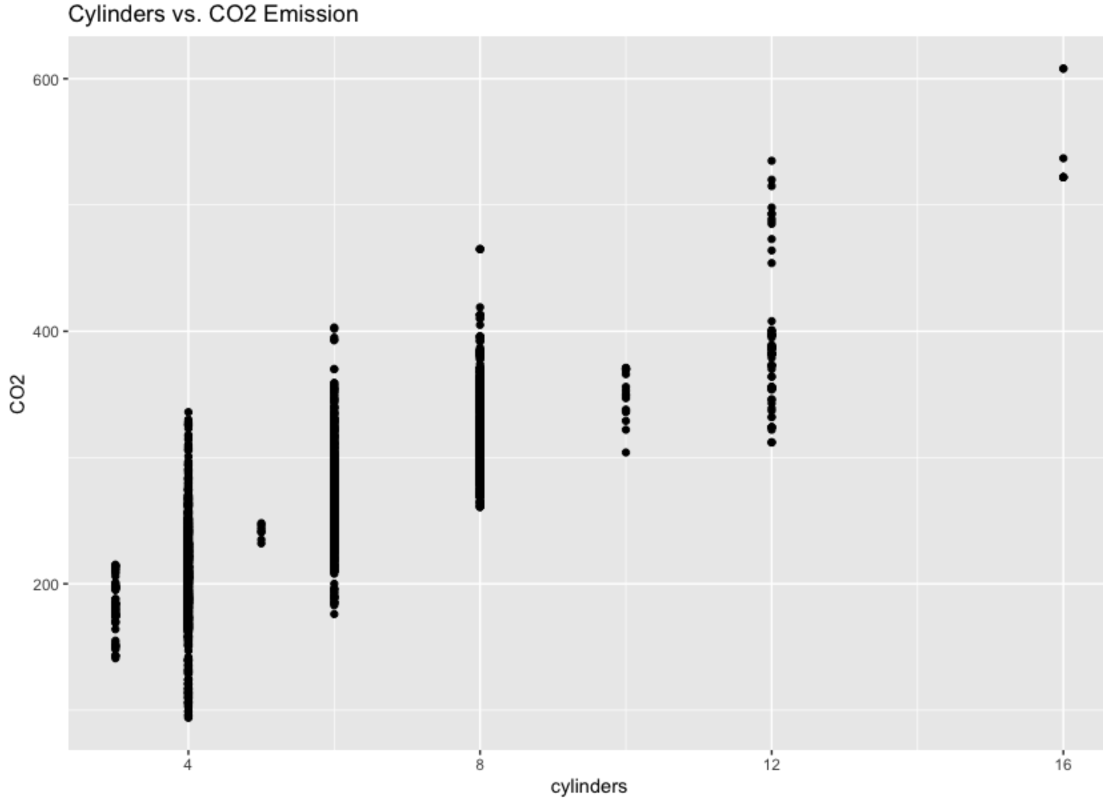
descriptive statistics

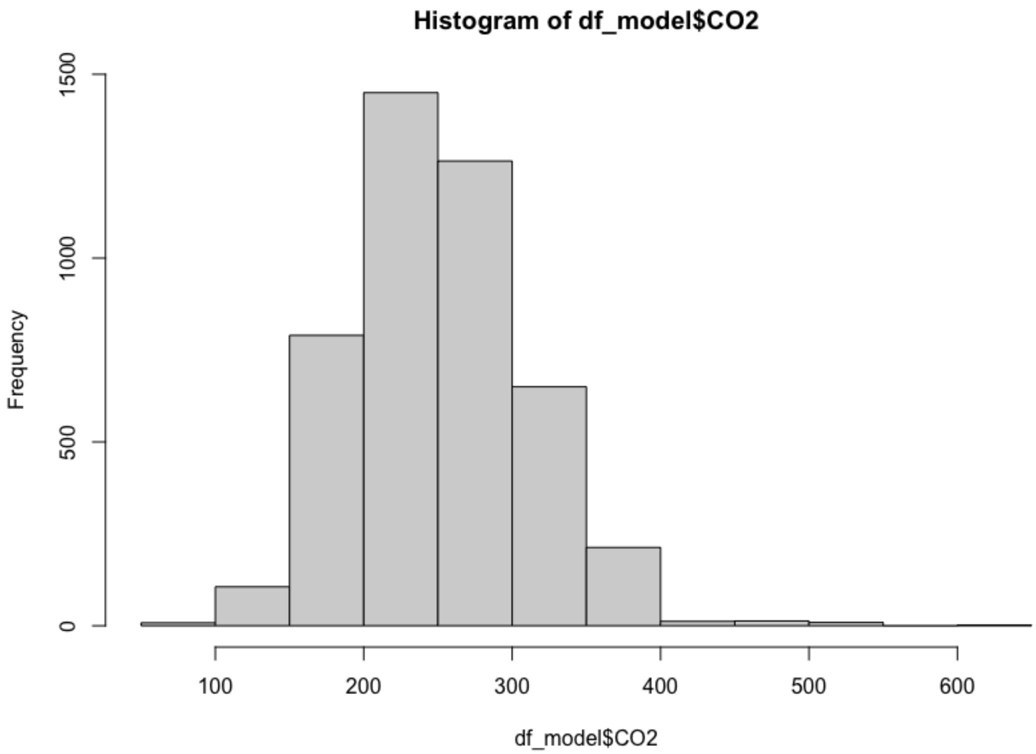
```
> summary(df_model)
      year          make              class            engine_size      cylinders      transmission        fuel_type        
 Min.   :2018   Length:4517        Length:4517        Min.   :1.000   Min.   : 3.000   Length:4517        Length:4517       
 1st Qu.:2019   Class :character   Class :character   1st Qu.:2.000   1st Qu.: 4.000   Class :character   Class :character  
 Median :2020   Mode  :character   Mode  :character   Median :3.000   Median : 6.000   Mode  :character   Mode  :character  
 Mean   :2020                                         Mean   :3.097   Mean   : 5.549                                        
 3rd Qu.:2021                                         3rd Qu.:3.600   3rd Qu.: 6.000                                        
 Max.   :2022                                         Max.   :8.000   Max.   :16.000                                        
   fuel_city        fuel_hwy           CO2         model_type       
 Min.   : 4.00   Min.   : 3.900   Min.   : 94.0   Length:4517       
 1st Qu.:10.00   1st Qu.: 7.600   1st Qu.:209.0   Class :character  
 Median :11.90   Median : 8.800   Median :247.0   Mode  :character  
 Mean   :12.27   Mean   : 9.075   Mean   :251.5                     
 3rd Qu.:14.30   3rd Qu.:10.300   3rd Qu.:289.0                     
 Max.   :30.30   Max.   :20.900   Max.   :608.0                     
```

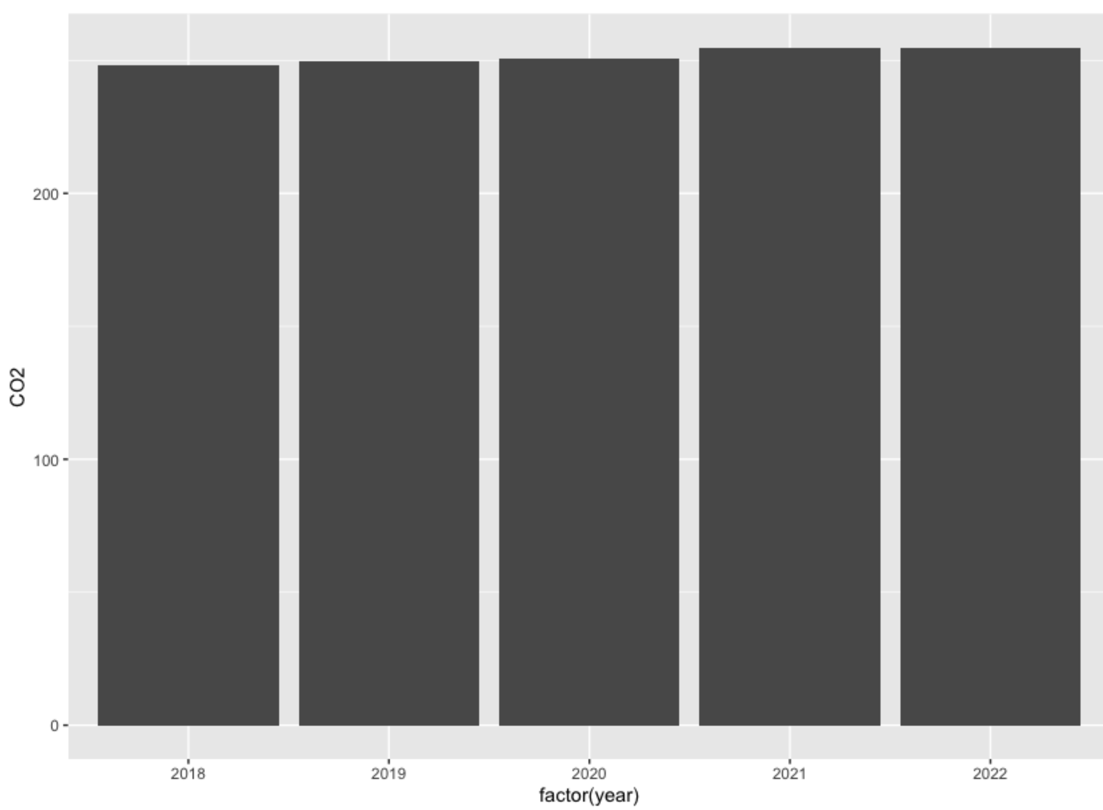
Engine Size vs. CO2 Emission



Cylinders vs. CO2 Emission




Histogram of df_model$CO2
Frequency
0
500
1000
1500
100
200
300
400
500
600
df_model$CO2


CO2
0
100
200
2018
2019
2020
2021
2022
factor(year)

Counts of each category

```
> sapply(df_model, n_distinct)
       year        make       class  engine_size    cylinders transmission    fuel_type    fuel_city     fuel_hwy          CO2
          5          39          15           44            8           26            4          192          131          298
 model_type
          7
> unique(df_model$make)
 [1] "Acura"         "Alfa Romeo"    "Aston Martin"  "Audi"          "Bentley"       "BMW"           "Bugatti"       "Buick"
 [9] "Cadillac"      "Chevrolet"     "Chrysler"      "Dodge"         "FIAT"          "Ford"          "Genesis"       "GMC"
[17] "Honda"         "Hyundai"       "Infiniti"      "Jaguar"        "Jeep"          "Kia"           "Lamborghini"   "Land Rover"
[25] "Lexus"         "Lincoln"       "Maserati"      "Mazda"         "Mercedes-Benz" "MINI"          "Mitsubishi"    "Nissan"
[33] "Porsche"       "Ram"           "Rolls-Royce"   "Subaru"        "Toyota"        "Volkswagen"    "Volvo"
> unique(df_model$class)
 [1] "Compact"                 "SUV: Small"              "Mid-size"                "Minicompact"             "SUV: Standard"
 [6] "Two-seater"              "Subcompact"              "Station wagon: Small"    "Station wagon: Mid-size" "Full-size"
[11] "Pickup truck: Small"     "Pickup truck: Standard"  "Minivan"                 "Special purpose vehicle" "Van: Passenger"
> unique(df_model$model_type)
[1] "Other" "AWD"   "4WD"   "FFV"   "LWB"   "SWB"   "EWB"
```

Part 3 - Feature Engineering

1. Create a column for the difference between the model year and the current year
2. Create a column for the difference between the city consumption and the hwy consumption
3. Remove columns city consumption and the hwy consumption
4. Create dummy variables

Part 4 – Fitting Model

1. Fit a model with all variables
2. Run a linear hypothesis on the variables we want to remove
3. Remove variables does not belong in the model

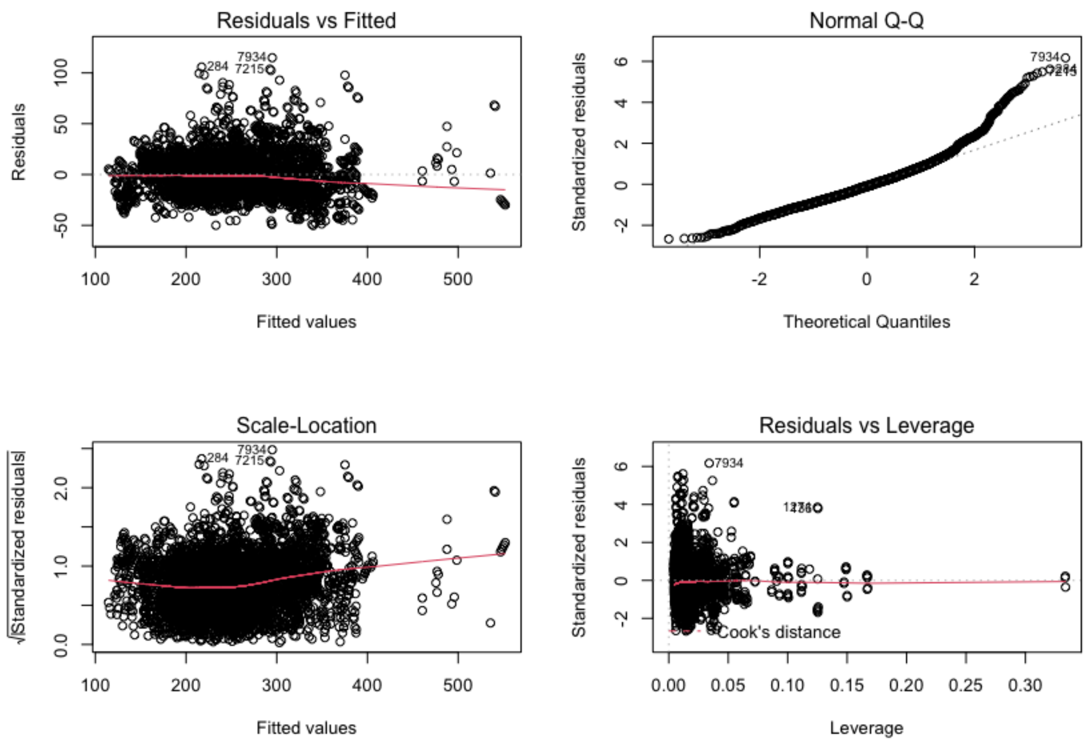
Output:

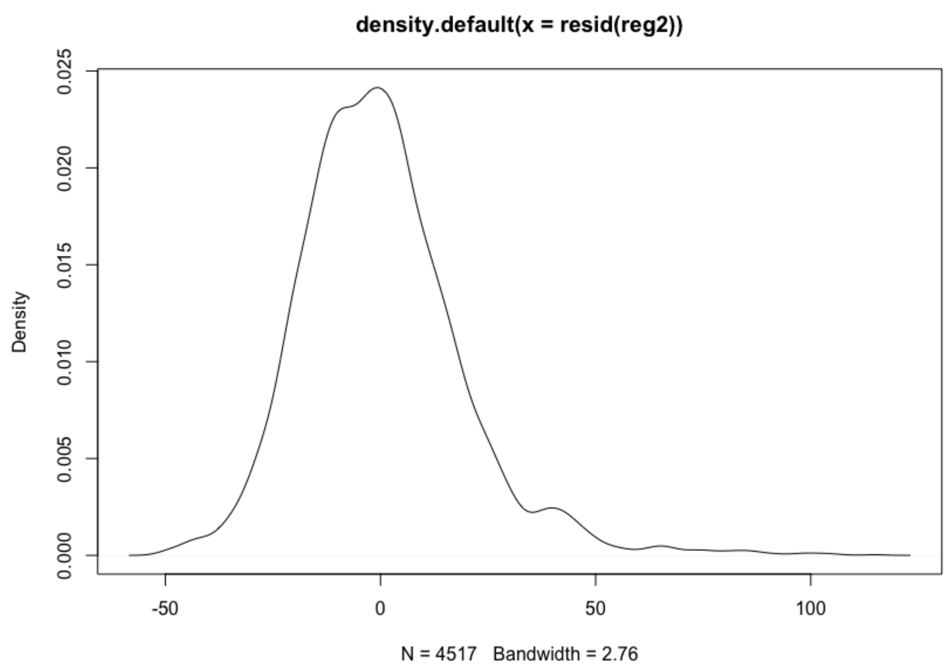
```
> summary(reg2)

Call:
lm(formula = CO2 ~ . - makeAcura - makeChrysler - makeInfiniti -
    makeJeep - makeLexus - makeMINI - makeRam - makeToyota -
    makeVolvo - classCompact - classFull.size - classMid.size -
    classMinicompact - transmissionA10 - transmissionAS9 - transmissionAV1 -
    transmissionAV7 - transmissionAV8 - transmissionM5 - fuel_typeD -
    fuel_typeX - model_type4WD, data = dataframe_dummyd)

Residuals:
    Min      1Q  Median      3Q     Max
-50.195 -12.391  -1.730   9.729 114.911

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                122.2677     2.2054  55.441  < 2e-16 ***
year                        -1.2970     0.2116  -6.128 9.64e-10 ***
makeAlfa.Romeo              12.1038     3.6881   3.282 0.001039 **
makeAston.Martin            11.5409     4.4125   2.616 0.008939 **
makeAudi                    12.1212     1.9012   6.376 2.01e-10 ***
makeBentley                 13.3195     3.9908   3.338 0.000852 ***
makeBMW                      9.1928     1.7410   5.280 1.35e-07 ***
makeBugatti                 97.1521     7.8212  12.422  < 2e-16 ***
makeBuick                    9.5306     2.6821   3.553 0.000384 ***
makeCadillac                11.9016     2.0999   5.668 1.54e-08 ***
makeChevrolet                7.3072     1.3619   5.365 8.50e-08 ***
makeDodge                    7.5282     2.2734   3.311 0.000936 ***
makeFIAT                    11.4210     4.4462   2.569 0.010240 *
makeFord                    20.0463     1.4884  13.468  < 2e-16 ***
makeGenesis                 20.7007     3.3311   6.214 5.63e-10 ***
makeGMC                      6.7021     1.7094   3.921 8.97e-05 ***
makeHonda                    6.9252     1.8549   3.734 0.000191 ***
makeHyundai                  6.9257     1.8470   3.750 0.000179 ***
makeJaguar                  11.5546     2.6822   4.308 1.68e-05 ***
makeKia                     11.6387     1.9553   5.952 2.85e-09 ***
makeLamborghini             58.3774     4.4111  13.234  < 2e-16 ***
makeLand.Rover              25.9350     2.9444   8.808  < 2e-16 ***
makeLincoln                 24.6477     2.8333   8.699  < 2e-16 ***
makeMaserati                43.9045     3.1803  13.805  < 2e-16 ***
makeMazda                  -10.0739     2.0699  -4.867 1.17e-06 ***
makeMercedes.Benz           15.9570     1.9041   8.380  < 2e-16 ***
makeMitsubishi              10.9969     3.3346   3.298 0.000982 ***
makeNissan                  11.2872     1.8448   6.118 1.03e-09 ***
makePorsche                 27.9653     1.8742  14.921  < 2e-16 ***
makeRolls.Royce             24.4630     4.6292   5.285 1.32e-07 ***
makeSubaru                  12.9767     2.3427   5.539 3.21e-08 ***
makeVolkswagen              15.1370     2.1648   6.992 3.11e-12 ***
classMinivan                26.0527     3.1787   8.196 3.22e-16 ***
classPickup.truck..Small    47.5210     2.2072  21.530  < 2e-16 ***
classPickup.truck..Standard 39.9537     1.5356  26.017  < 2e-16 ***
classSpecial.purpose.vehicle 40.9399    2.9176  14.032  < 2e-16 ***
classStation.wagon..Mid.size  8.0188    3.0932   2.592 0.009561 **
classStation.wagon..Small    9.3236     1.7855   5.222 1.85e-07 ***
classSubcompact              3.4128     1.2175   2.803 0.005083 **
classSUV..Small             24.0376     0.9418  25.524  < 2e-16 ***
```

```
classSUV..Standard          35.1310     1.2062  29.124  < 2e-16 ***
classTwo.seater              7.8894     1.6470   4.790 1.72e-06 ***
classVan..Passenger         87.2254     5.6617  15.406  < 2e-16 ***
engine_size                 10.7954     0.8241  13.100  < 2e-16 ***
cylinders                    6.1038     0.5865  10.407  < 2e-16 ***
transmissionA4              25.5283     7.3552   3.471 0.000524 ***
transmissionA5              27.1122     4.7556   5.701 1.27e-08 ***
transmissionA6              12.8573     1.8220   7.057 1.97e-12 ***
transmissionA7              21.5162     3.7698   5.708 1.22e-08 ***
transmissionA8               6.9919     1.6469   4.246 2.22e-05 ***
transmissionA9               6.5778     1.7875   3.680 0.000236 ***
transmissionAM6            -16.3459     2.9295  -5.580 2.55e-08 ***
transmissionAM7             11.0357     1.8682   5.907 3.74e-09 ***
transmissionAM8             15.7621     2.3535   6.697 2.39e-11 ***
transmissionAM9             63.9679    11.2342   5.694 1.32e-08 ***
transmissionAS10             8.4616     1.7742   4.769 1.91e-06 ***
transmissionAS5             50.0710     7.9265   6.317 2.93e-10 ***
transmissionAS6             14.9411     1.5854   9.424  < 2e-16 ***
transmissionAS7             17.2740     2.8619   6.036 1.71e-09 ***
transmissionAS8              6.5423     1.4612   4.477 7.74e-06 ***
transmissionAV             -11.1652     1.7921  -6.230 5.09e-10 ***
transmissionAV10           -10.4972     4.0478  -2.593 0.009536 **
transmissionAV6            -14.4332     3.0486  -4.734 2.27e-06 ***
transmissionM6              14.5115     1.5308   9.480  < 2e-16 ***
transmissionM7              12.9299     3.1187   4.146 3.45e-05 ***
fuel_typeE                 -39.6061     2.1616 -18.323  < 2e-16 ***
fuel_typeZ                  13.9131     0.9898  14.057  < 2e-16 ***
model_typeAWD              -16.5983     1.2615 -13.158  < 2e-16 ***
model_typeEWB              -26.3623     7.3349  -3.594 0.000329 ***
model_typeFFV              -25.8030     2.0752 -12.434  < 2e-16 ***
model_typeLWB              -30.3553     4.4653  -6.798 1.20e-11 ***
model_typeOther            -23.7525     1.1079 -21.439  < 2e-16 ***
model_typeSWB              -37.5259     6.5348  -5.743 9.95e-09 ***
fuel_d                      13.6662     0.4030  33.908  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.95 on 4443 degrees of freedom
Multiple R-squared:  0.9021,    Adjusted R-squared:  0.9005
F-statistic: 560.6 on 73 and 4443 DF,  p-value: < 2.2e-16
```

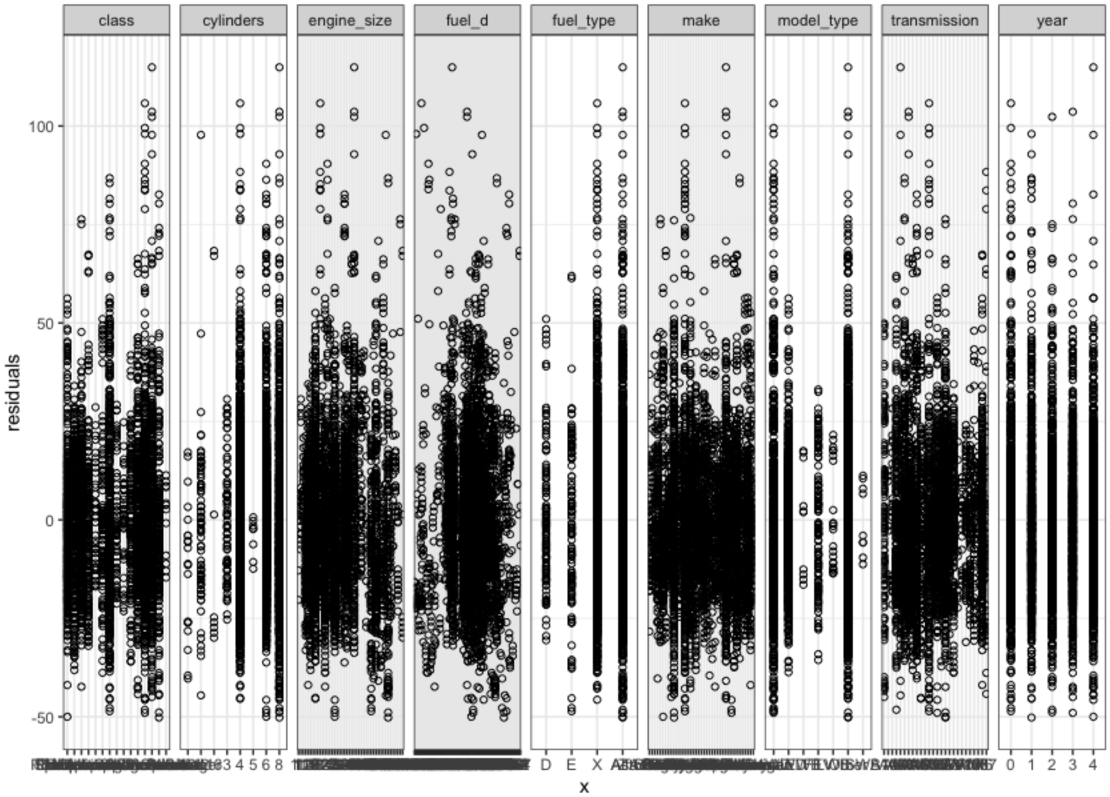
Part 5 – Testing and Tuning Model

1. Plot to test model





2. Test for heteroskedasticity



There is heteroskedasticity in our current model.

```
> ncvTest(reg2) #rejected the null hypothesis of no heteroskedasticity
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 245.3856, Df = 1, p = < 2.22e-16
```

3. Correct the model using HCCME

Part of the output

transmissionAS5	50.071	3.0535	16.398	1.020e-58	44.0846	56.0574	4443
transmissionAS6	14.941	1.5452	9.669	6.676e-22	11.9117	17.9704	4443
transmissionAS7	17.274	2.2079	7.824	6.365e-15	12.9454	21.6026	4443
transmissionAS8	6.542	1.3481	4.853	1.258e-06	3.8993	9.1852	4443
transmissionAV	-11.165	1.5686	-7.118	1.273e-12	-14.2405	-8.0899	4443
transmissionAV10	-10.497	2.1368	-4.913	9.311e-07	-14.6864	-6.3080	4443
transmissionAV6	-14.433	2.6467	-5.453	5.214e-08	-19.6221	-9.2443	4443
transmissionM6	14.512	1.4798	9.807	1.780e-22	11.6104	17.4126	4443
transmissionM7	12.930	5.1269	2.522	1.170e-02	2.8787	22.9812	4443
fuel_typeE	-39.606	2.5165	-15.739	2.358e-54	-44.5397	-34.6726	4443
fuel_typeZ	13.913	0.9987	13.932	3.259e-43	11.9552	15.8710	4443
model_typeAWD	-16.598	1.3473	-12.320	2.552e-34	-19.2396	-13.9570	4443
model_typeEWB	-26.362	5.5943	-4.712	2.523e-06	-37.3298	-15.3947	4443
model_typeFFV	-25.803	2.1207	-12.167	1.583e-33	-29.9607	-21.6453	4443
model_typeLWB	-30.355	3.4005	-8.927	6.297e-19	-37.0219	-23.6887	4443
model_typeOther	-23.753	1.2632	-18.804	5.809e-76	-26.2290	-21.2761	4443
model_typeSWB	-37.526	3.4734	-10.804	7.117e-27	-44.3356	-30.7162	4443
fuel_d	13.666	0.6301	21.689	3.081e-99	12.4309	14.9015	4443

```
Multiple R-squared:  0.9021 ,   Adjusted R-squared:  0.9005
F-statistic:   654 on 73 and 4443 DF,  p-value: < 2.2e-16
```

Part 6 – Make Predictions

1. Split data for train set (60%) and test set (40%)
2. Make prediction

Output

```
> summary(pred)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  123.7   210.9   245.3   251.2   285.7   551.9
> # test & train comparisons
> data.frame( R2 = R2(pred, test$CO2),
+             RMSE = RMSE(pred, test$CO2),
+             MAE = MAE(pred, test$CO2))
         R2     RMSE      MAE
1 0.9018035 19.07063 14.13455
```

Part 7 – Chow Test: does a two-seater respond more to change in Engine Size? (optional)

```
  Res.Df     RSS Df Sum of Sq     F    Pr(>F)
1   4444 1603897
2   4442 1595442  2    8455.1 11.77 7.976e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```