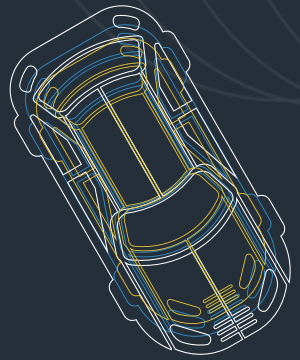
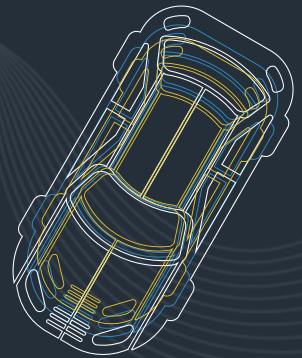


DANA 4810

CO2 Emissions in cars

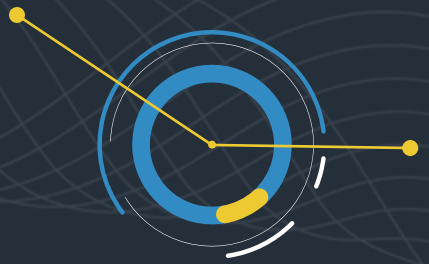


Objective

1. 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.

https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64#wb-auto-6

Stages



01

EDA

02

Stepwise

03

All possible subsets

04

Reduced model

05

Results

06

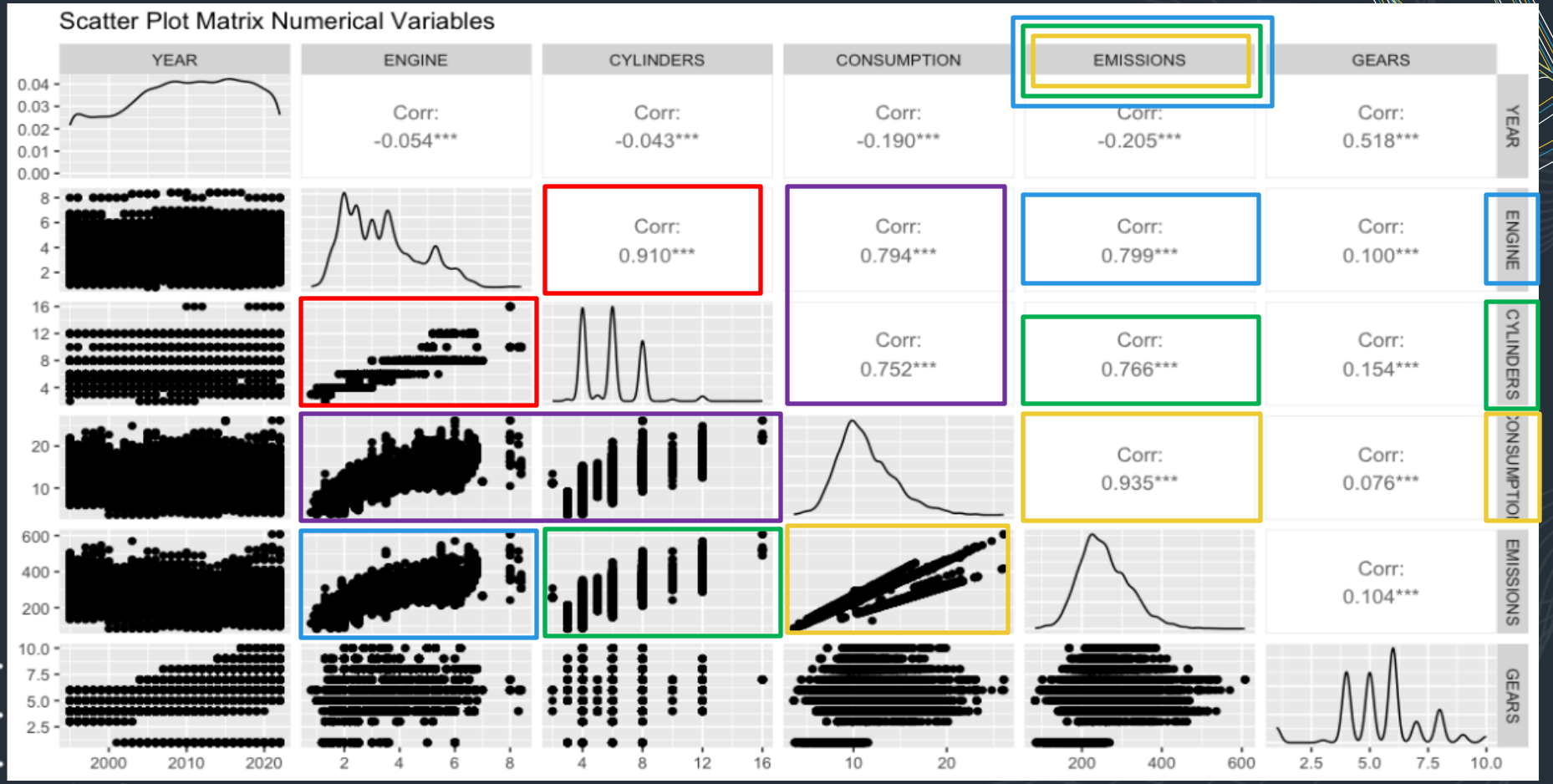
Conclusion

EDA

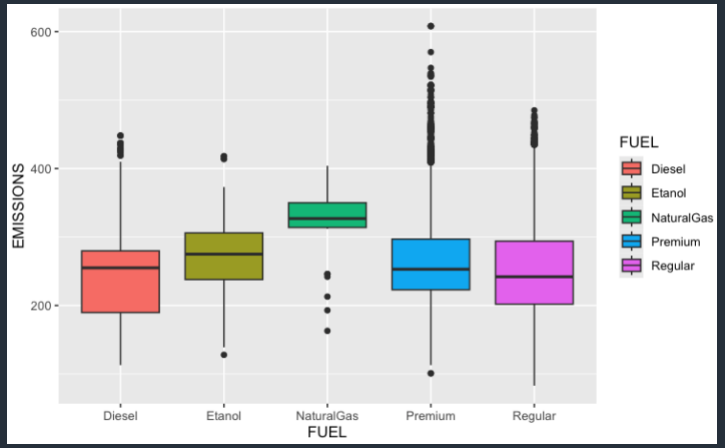
No	Pr. 1	Pr. 2	Pr. 3	Pr. 4	Pr. 5	Pr. 6	Pr. 7	Pr. 8	Pr. 9	Pr. 10	Resp. 1
Variable	YEAR	BRAND	MODEL	CLASS	ENGINE	CYLINDERS	TRANSMISSION	GEARS	FUEL	COMSUMPTION	EMISSIONS
Type	Numerical discrete	Categorical nominal	Categorical nominal	Categorical nominal	Numerical discrete	Numerical discrete	Categorical nominal	Numerical discrete	Categorical nominal	Numerical - continuos	Numerical - continuos
Unique Values	28	55	4185	17	64	9	5	5	5		
Range	1995 to 2022	-	-	-	0.8 to 8.4	2 to 16	-	1 to 10	-	8.7 to 26.1	128 to 418
Units / Categories		-	-	-			1. Automatic 2. Automated Manual 3. Automatic Shift 4. Continuous Variable 5. Manual		1. Diesel 2. Ethanol 3. Natural Gas 4. Premium 5. Regular	(Liter/100Km)	(g/Km)

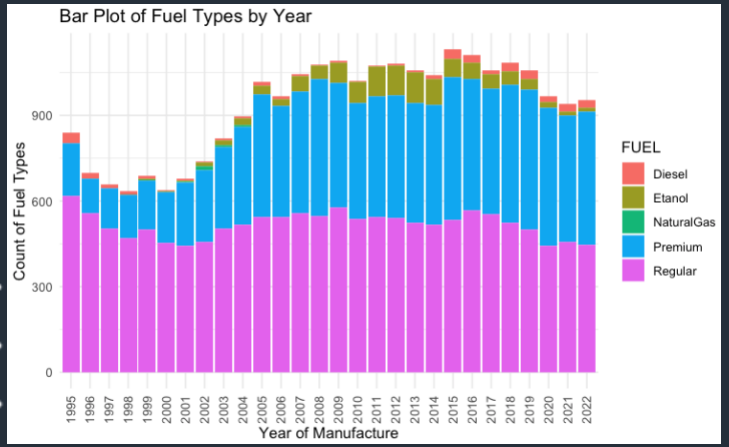
Train data Length:	20,860	80%
Test data Length:	5,215	20%
Total:	26,075	100%

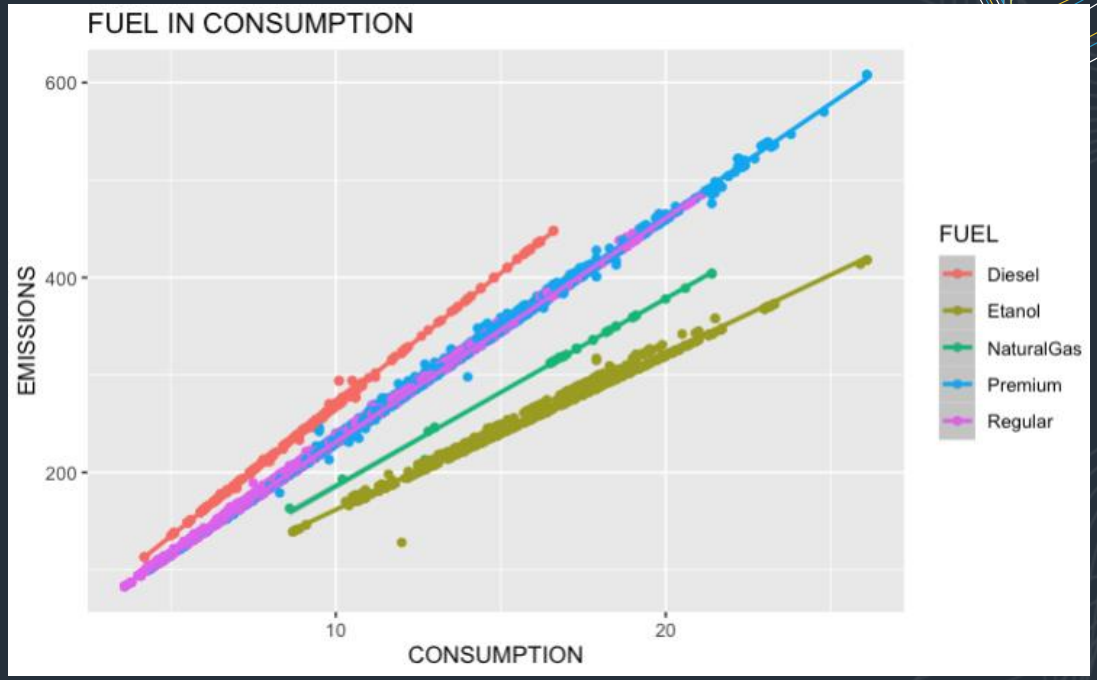
EDA-Numerical



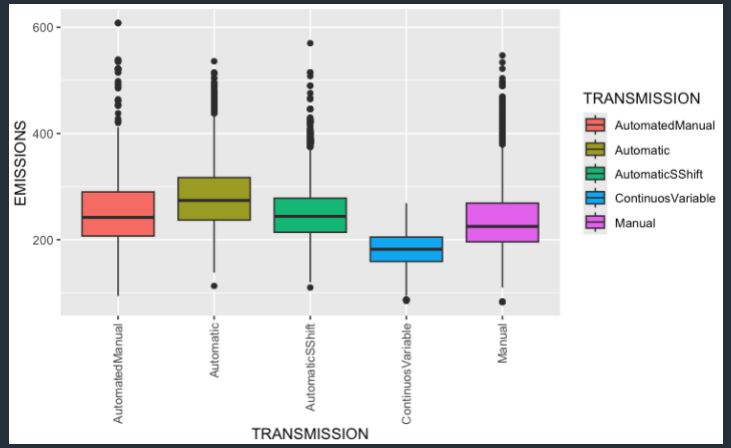
EDA-Categorical: Fuel

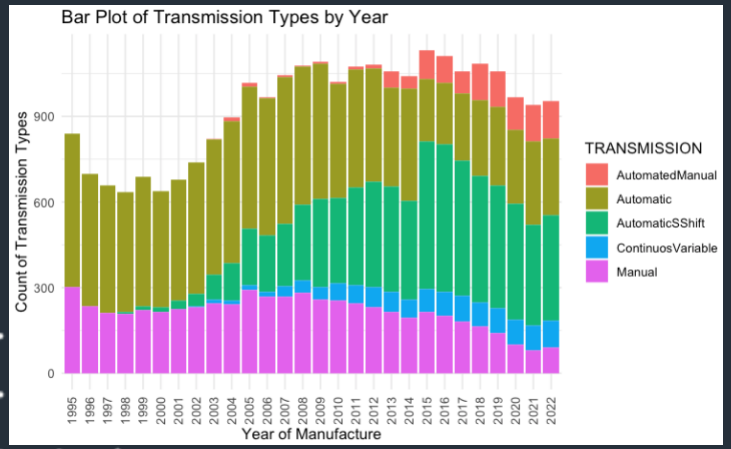


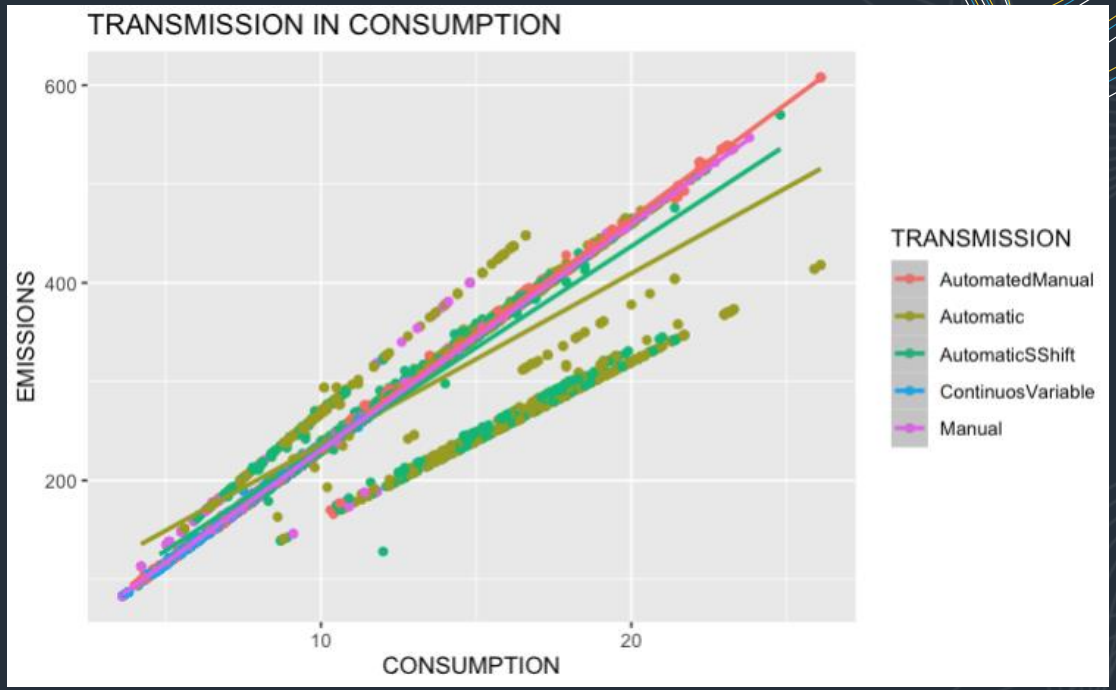




EDA-Categorical: Transmission



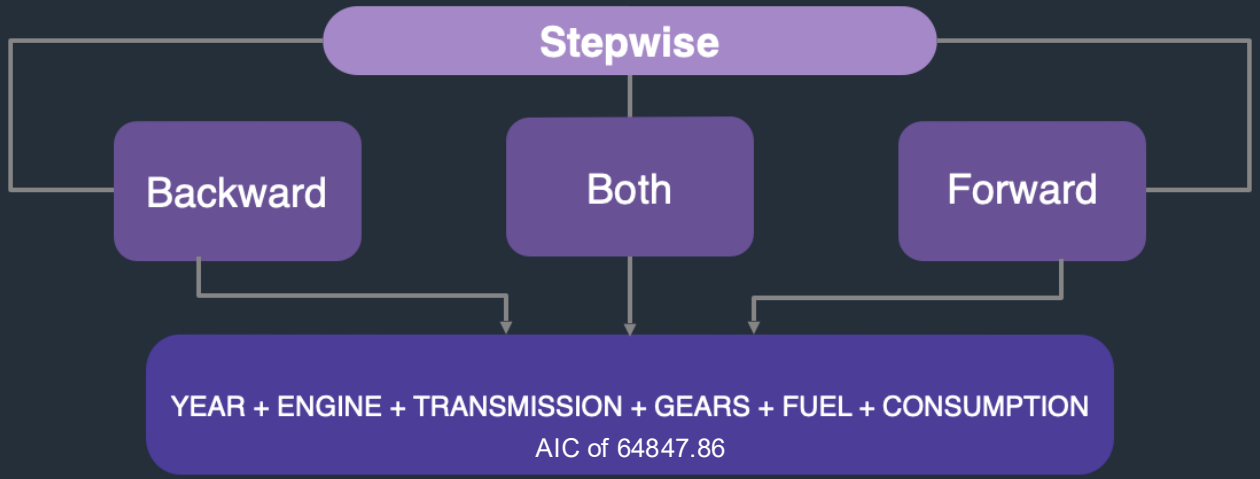




All Possible Subsets

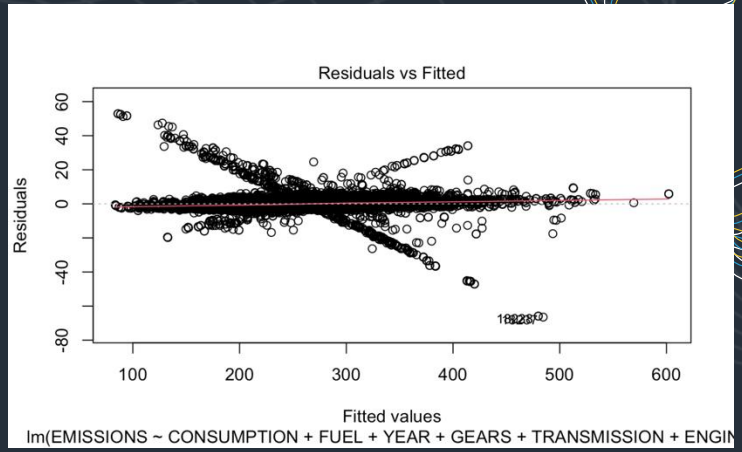
	Consumption	Cylinders	Engine	Gears	Years	Fuel E	Fuel NG	Fuel P	Fuel R	Trans A	Trans AS	Trans CV	Trans M
1	*												
2	*					*							
3	*					*	*						
4	*					*		*	*				
5	*					*	*	*	*				
6	*				*	*	*	*	*				*
7	*			*	*	*	*	*	*				*
8	*			*	*	*	*	*	*			*	*
9	*			*	*	*	*	*	*			*	*
10	*		*	*	*	*	*	*	*			*	*
11	*			*	*	*	*	*	*	*	*	*	*
12	*		*	*	*	*	*	*	*	*	*	*	*
13	*	*	*	*	*	*	*	*	*	*	*	*	*

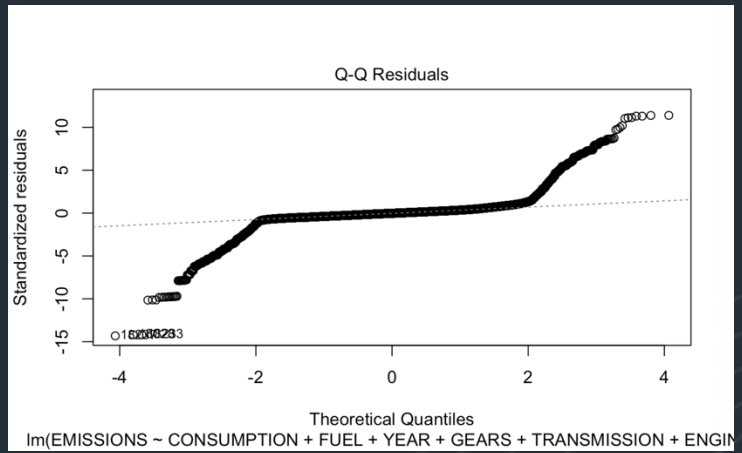
Stepwise


Stepwise
Backward
Both
Forward
YEAR + ENGINE + TRANSMISSION + GEARS + FUEL + CONSUMPTION
AIC of 64847.86

Model Without Interaction

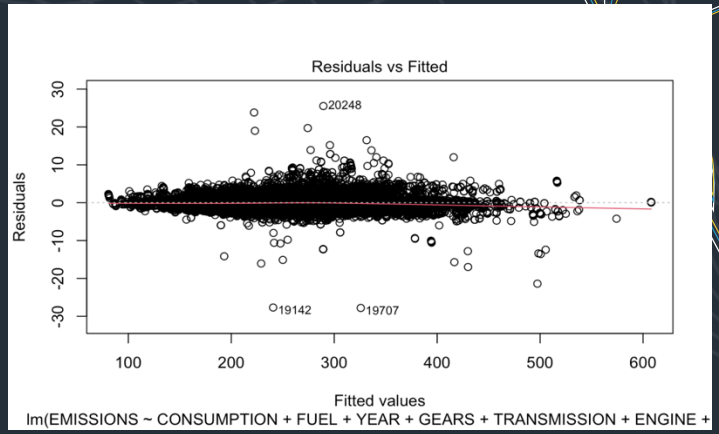
```
lm(formula = EMISSIONS ~ CONSUMPTION + FUEL + YEAR + GEARS +
    TRANSMISSION + ENGINE, data = co2_train)

Residuals:
    Min      1Q  Median      3Q     Max
-66.466  -1.263  -0.129   1.010  52.920

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   -1.760e+02  1.370e+01  -12.842  < 2e-16 ***
CONSUMPTION                    2.290e+01  2.121e-02 1079.869  < 2e-16 ***
FUELEtanol                    -1.505e+02  3.328e-01 -452.104  < 2e-16 ***
FUELNaturalGas                -1.027e+02  8.577e-01 -119.732  < 2e-16 ***
FUELPremium                   -3.447e+01  2.705e-01 -127.417  < 2e-16 ***
FUELRegular                   -3.397e+01  2.701e-01 -125.755  < 2e-16 ***
YEAR                           1.044e-01  6.866e-03   15.203  < 2e-16 ***
GEARS                          6.526e-01  3.623e-02   18.010  < 2e-16 ***
TRANSMISSIONAutomatic         -5.499e-01  1.801e-01   -3.053 0.002269 **
TRANSMISSIONAutomaticSShift   -6.007e-01  1.745e-01   -3.442 0.000579 ***
TRANSMISSIONContinuosVariable  1.637e+00  2.933e-01    5.580 2.44e-08 ***
TRANSMISSIONManual            -9.164e-01  1.824e-01   -5.025 5.07e-07 ***
ENGINE                        -9.683e-02  4.252e-02   -2.277 0.022785 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.647 on 20847 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9946,    Adjusted R-squared:  0.9945
F-statistic: 3.171e+05 on 12 and 20847 DF,  p-value: < 2.2e-16
```

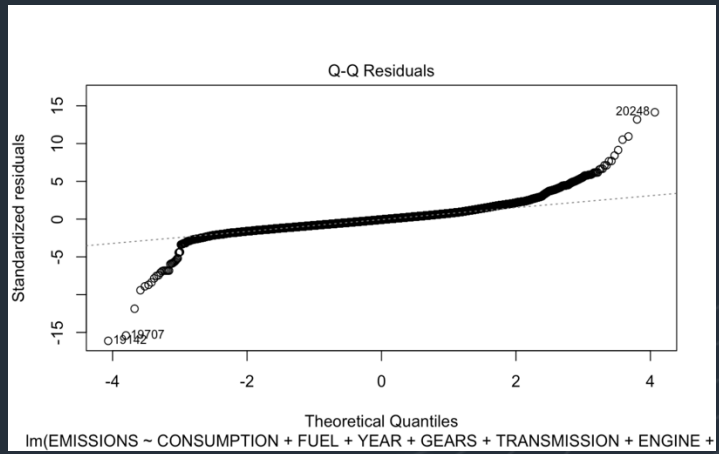




Model With Interaction

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

Residuals:
     Min       1Q   Median       3Q      Max
-27.8000  -1.0654  -0.1144   0.8556  25.5270

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   -3.649e+02  5.366e+00  -68.001  < 2e-16 ***
CONSUMPTION                    2.731e+01  3.664e-02  745.552  < 2e-16 ***
FUELEtanol                     1.796e+00  5.045e-01    3.561  0.00037 ***
FUELNaturalGas                -3.525e+00  1.923e+00   -1.833  0.06684 .
FUELPremium                    7.998e-01  3.580e-01    2.234  0.02547 *
FUELRegular                    1.545e+00  3.539e-01    4.365 1.28e-05 ***
YEAR                           1.795e-01  2.683e-03   66.908  < 2e-16 ***
GEARS                          3.585e-01  1.411e-02   25.412  < 2e-16 ***
TRANSMISSIONAutomatic         -2.992e-01  7.011e-02   -4.268 1.98e-05 ***
TRANSMISSIONAutomaticSShift   -5.441e-01  6.782e-02   -8.023 1.08e-15 ***
TRANSMISSIONContinuosVariable  1.283e+00  1.140e-01   11.252  < 2e-16 ***
TRANSMISSIONManual            -4.416e-01  7.087e-02   -6.231 4.73e-10 ***
ENGINE                        -4.993e-01  1.661e-02  -30.071  < 2e-16 ***
CONSUMPTION:FUELEtanol        -1.104e+01  4.192e-02 -263.476  < 2e-16 ***
CONSUMPTION:FUELNaturalGas    -7.839e+00  1.166e-01  -67.240  < 2e-16 ***
CONSUMPTION:FUELPremium       -3.924e+00  3.670e-02 -106.900  < 2e-16 ***
CONSUMPTION:FUELRegular       -3.952e+00  3.641e-02 -108.533  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.805 on 20843 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9992,    Adjusted R-squared:  0.9992
F-statistic: 1.583e+06 on 16 and 20843 DF,  p-value: < 2.2e-16
```





Reduced Model With Interaction

Based on the observation of the results in the “All possible subsets” and the EDA, a reduced model is appropriate:

EMISSIONS ~ CONSUMPTION + FUEL + CONSUMPTION * FUEL

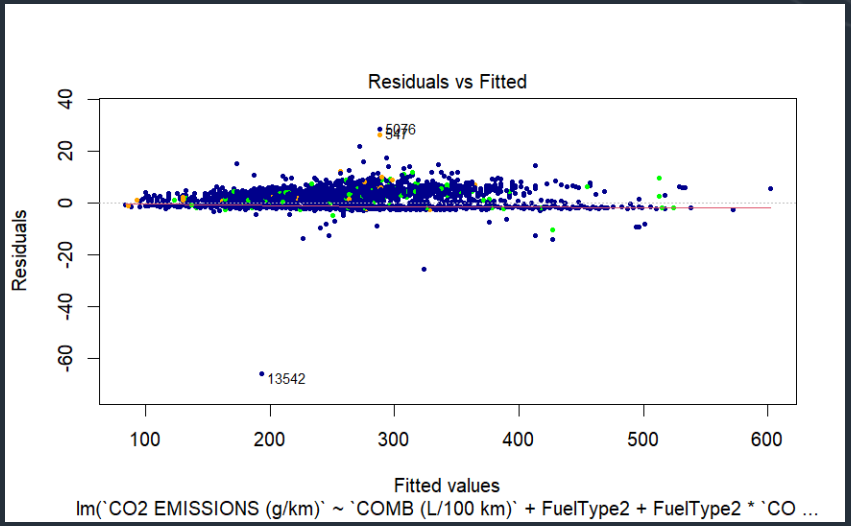
```
lm(formula = `CO2 EMISSIONS (g/km)` ~ `COMB (L/100 km)` + FuelType2 +
    FuelType2 * `COMB (L/100 km)`, data = train2)

Residuals:
    Min      1Q  Median      3Q     Max
-65.704  -1.458  -1.097   0.997  28.539

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                     -0.39809    0.47463   -0.839    0.402
`COMB (L/100 km)`               26.98672    0.04831  558.632   <2e-16 ***
FuelType2E                       1.37627    0.68314    2.015    0.044 *
FuelType2G                       1.15665    0.48034    2.408    0.016 *
FuelType2N                       1.46427    2.78467    0.526    0.599
`COMB (L/100 km)`:FuelType2E -10.92621    0.05630 -194.072   <2e-16 ***
`COMB (L/100 km)`:FuelType2G  -3.93953    0.04874  -80.828   <2e-16 ***
`COMB (L/100 km)`:FuelType2N  -8.14635    0.16427  -49.590   <2e-16 ***
```

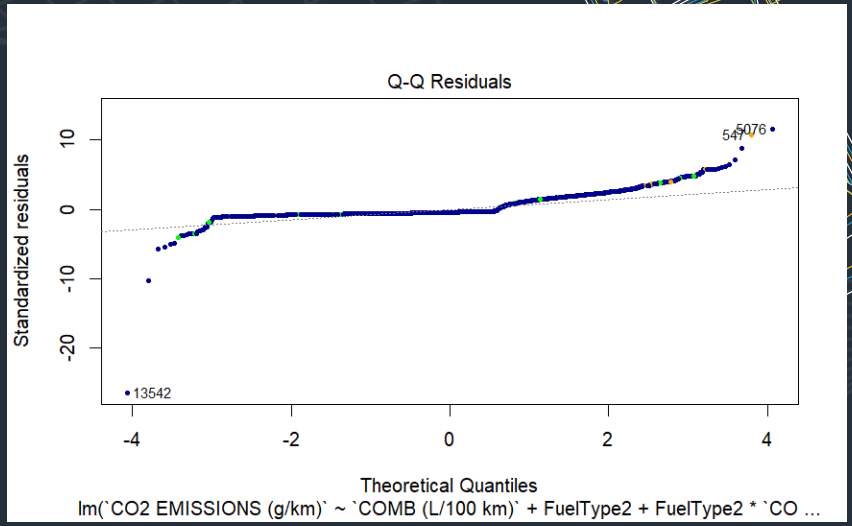
Accuracy Test In Training Data Results

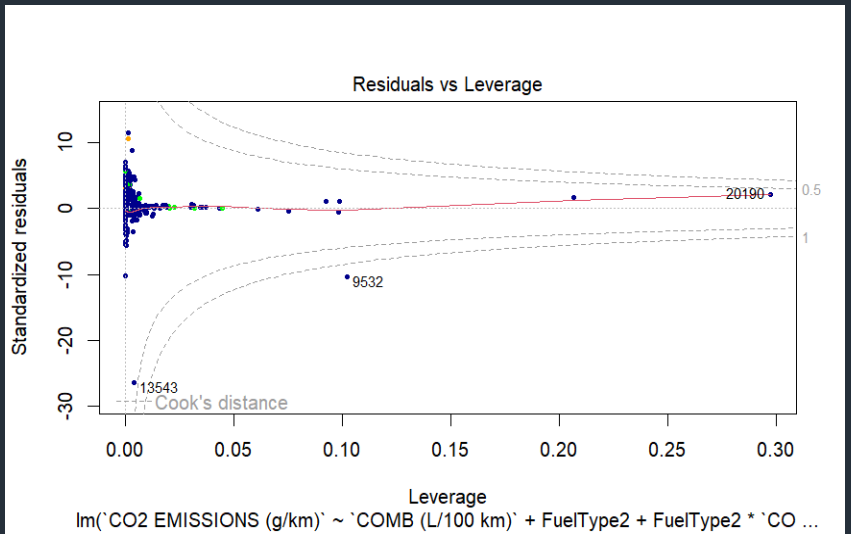
```
"R Squared:  0.998445138435789"
"Adjusted R Squared:  0.998444616444952"
"AIC Score:  97238.827549705"
"MSE Score:  6.19235906016588"
```

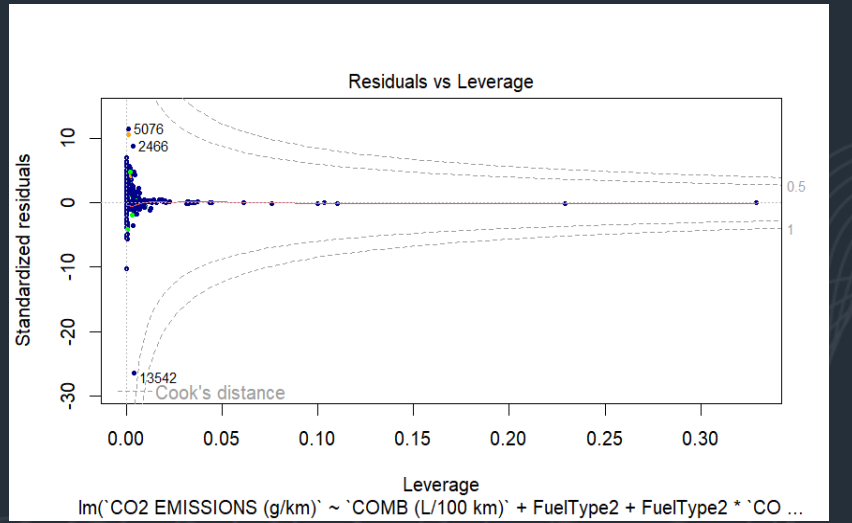
Tested Against Test Data:

```
MSE on test data: 6.350973
RMSE on test data: 2.520114
R-squared on test data: 0.9983957
```


Residuals vs Fitted
Residuals
Fitted values
lm(`CO2 EMISSIONS (g/km)` ~ `COMB (L/100 km)` + FuelType2 + FuelType2 * `CO ...
5076
5077
13542

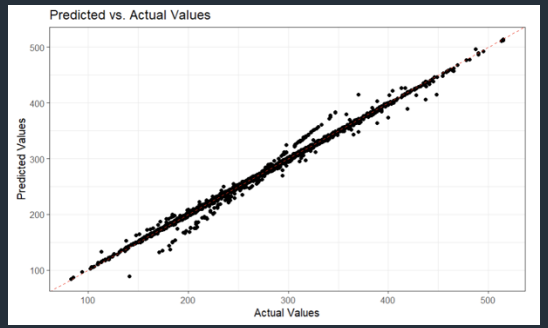

Q-Q Residuals
Standardized residuals
Theoretical Quantiles
lm(`CO2 EMISSIONS (g/km)` ~ `COMB (L/100 km)` + FuelType2 + FuelType2 * `CO ...
5076
5077
13542


Residuals vs Leverage
Standardized residuals
Leverage
lm(`CO2 EMISSIONS (g/km)` ~ `COMB (L/100 km)` + FuelType2 + FuelType2 * `CO ...
Cook's distance
0.5
1
20190
9532
13543


Residuals vs Leverage
Standardized residuals
Leverage
lm(`CO2 EMISSIONS (g/km)` ~ `COMB (L/100 km)` + FuelType2 + FuelType2 * `CO ...
Cook's distance
0.5
1
5076
2466
13542

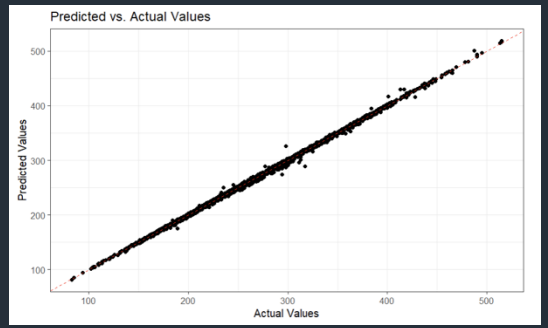
Results in test data

Stepwise model



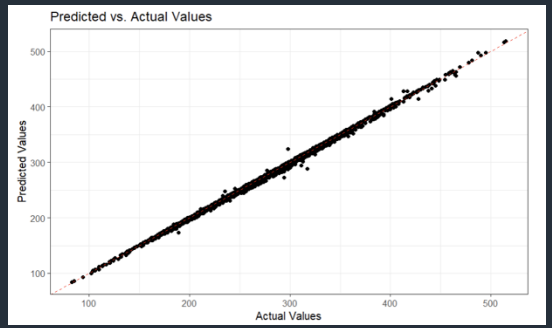
RMSE: 4.76

Stepwise model with interaction



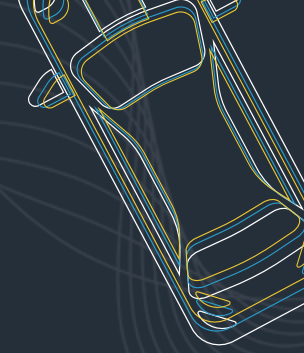
RMSE: 1.90

Reduced model with interaction

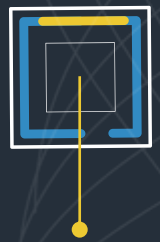


RMSE: 2.52

Conclusion



- Model with similar results, selected the simpler model (consumption and fuel type with interaction)
- Variables such as engine, cylinders and consumption are really correlated to emissions
- The type of fuel is the variable the determine how steep is the line of prediction



Thanks!

Do you have any questions?