JUDY's Carbon Crusaders:

CO_2 Emission by Vehicles

Jonathan Serrano, Umar Ali-Salaam, Daniel Li, Yaseen Mohammed

Motivation & Analysis Goal

Motivation:

From our understanding, the CO2 emissions of cars are a significant contributor to global warming, and as such, their impact on our world should not be ignored.

Analysis Goals:

The goal of our analysis is to determine which factors of a car, such as fuel consumption, the amount of cylinders, fuel type, and engine size, are best at modeling the amount of correlation to the CO2 emissions of each car.

Describing Our Dataset

- We found our dataset on Kaggle, however the origin of the data is from the Canadian Government official open data website.
- Since the data originates from a government institution, we believe that the data is reliable.
- The data contains 7385 observations of vehicles from a period of 7 years, 6282 of them being unique.
- There are 12 columns that describe certain features of the car, statistics of the fuel, and CO2 emissions.

Data Variables

Regressor Variables

Vehicle Class
(Discrete)

Engine Size
(Ranges from 0.9 to 8.4)

Transmission Type
(Discrete)

Fuel Type
(Discrete)

Fuel Consumption Hwy
(Ranges : 4-20.6) (L/100 km)

Fuel Consumption Comb
(Ranges : 4.1 - 26.1)(L/100 km)

Car Model
(Discrete)

Car Make
(Discrete)

Cylinders
(Discrete)

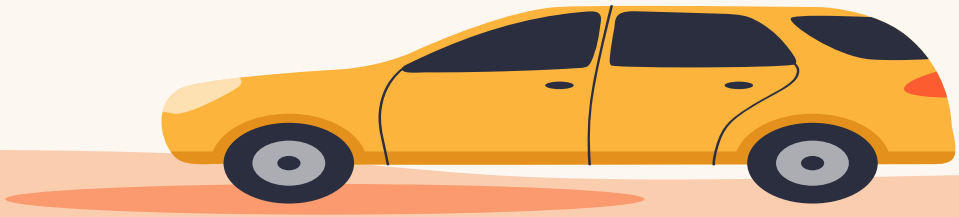
Fuel Consumption City
(Ranges : 4.2 - 30.6)(L/100 km)

Predictor Variable

CO2 Emissions (g/km)
(Ranges from 96 to 522)

Data Cleaning

- There was no NA values
- Removed unnecessary column (Fuel Consumption Comb (mpg))
- Added Dummy variable columns for discrete variables

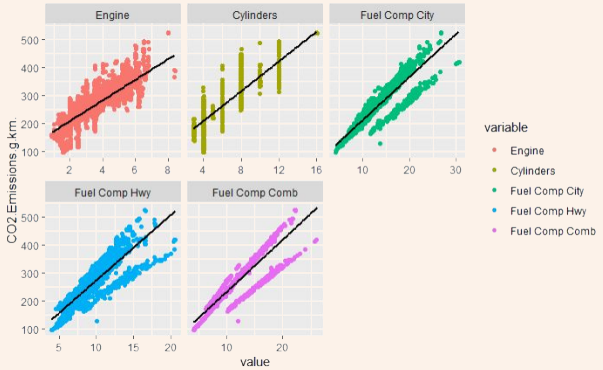


01

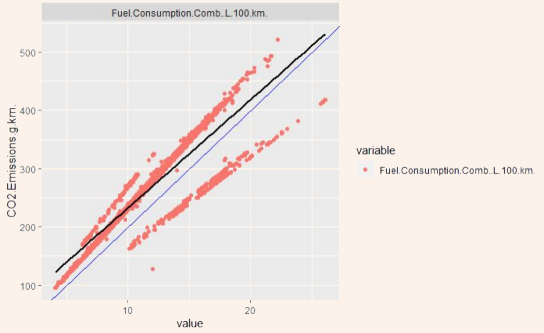
Data Exploration

Simplifying Data

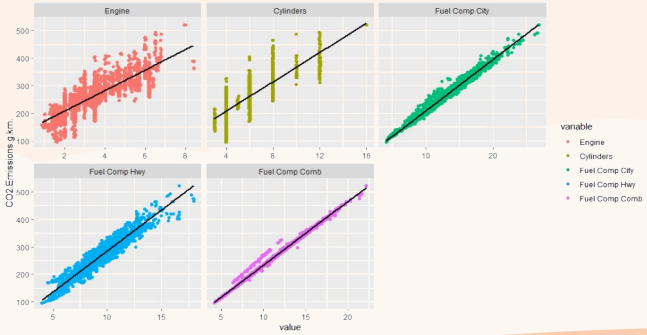
1. Original Correlation Plots



2. Separated Distributions



3. End Result



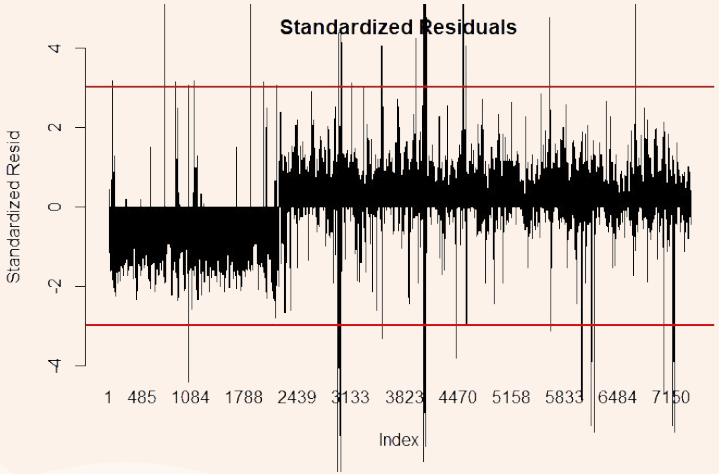
Correlation Matrix

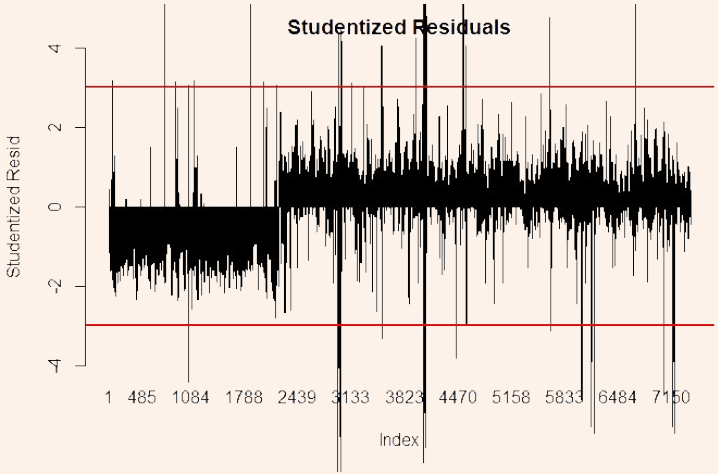
	Engine	Cylinders	Fuel Comp City	Fuel Comp Hwy	Fuel Comp Comb	CO2 Emissions
Engine	1.0000000	0.9259199	0.8616495	0.7956075	0.8518779	0.8514051
Cylinders	0.9259199	1.0000000	0.8468025	0.7641587	0.8310884	0.8321540
Fuel Comp City	0.8616495	0.8468025	1.0000000	0.9353228	0.9926066	0.9882668
Fuel Comp Hwy	0.7956075	0.7641587	0.9353228	1.0000000	0.9709317	0.9662925
Fuel Comp Comb	0.8518779	0.8310884	0.9926066	0.9709317	1.0000000	0.9953828
C02 Emissions	0.8514051	0.8321540	0.9882668	0.9662925	0.9953828	1.0000000

02

Residual Analysis

Residuals

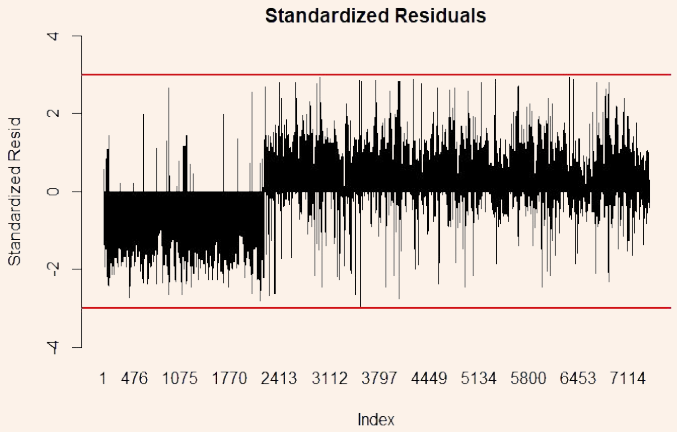


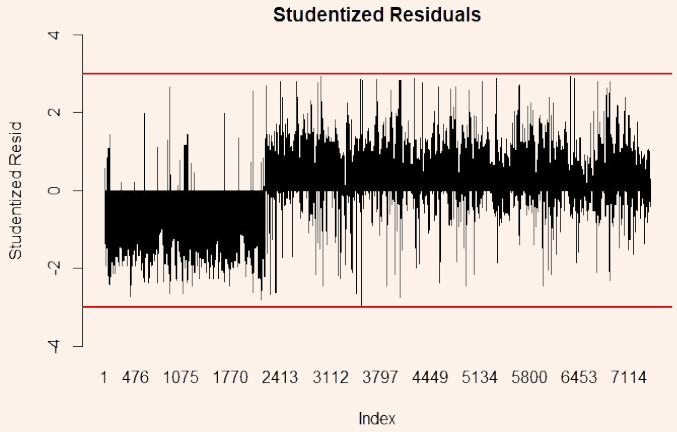


Standardized Residual

Studentized Residual

Residuals

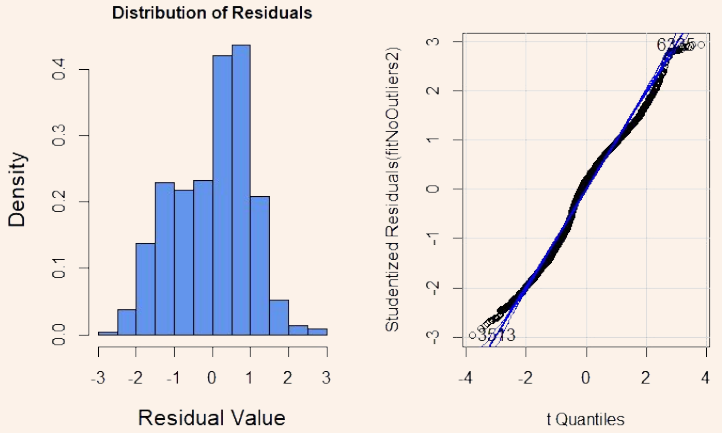




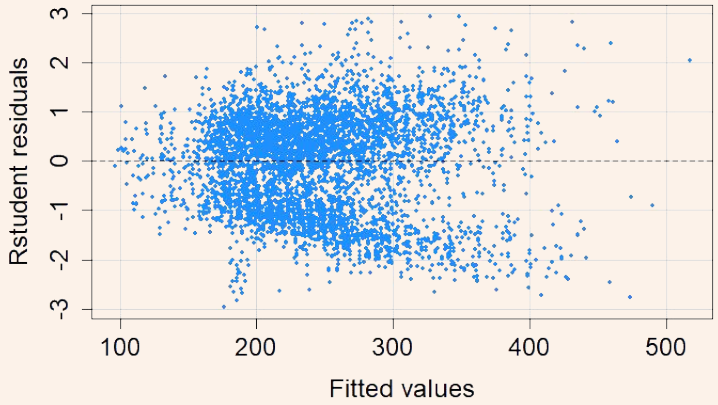
Standardized Residual

Studentized Residual

Residuals



Distribution of Residuals



Residual Plot

03

VIF Analysis

VIF Test

	GVIF
Engine.Size.L.	10.516585
Cylinders	9.638679
Fuel.Consumption.City..L.100.km.	1703.791684
Fuel.Consumption.Hwy..L.100.km.	438.338248
Fuel.Consumption.Comb..L.100.km.	3682.298415
Fuel.Type	2.620058
Transmission	13.712245
Vehicle.Class	17.165901

Fuel Consumption Per Road Type

VIF Test

	GVIF
Engine.Size.L.	4.312856
Fuel.Consumption.Comb..L.100.km.	5.188675
Fuel.Type	1.717297
Vehicle.Class	2.705236

	GVIF
Engine.Size.L.	4.300666
Fuel.Consumption.Comb..L.100.km.	4.541783
Fuel.Type	1.887630
Transmission	2.725601

Vehicle Class Vs. Transmission

VIF Test

	GVIF
Cylinders	4.753246
Fuel.Consumption.Comb..L.100.km.	6.156082
Fuel.Type	2.551260
Vehicle.Class	10.768431
Transmission	9.694901

	GVIF
Engine.Size.L.	5.116296
Fuel.Consumption.Comb..L.100.km.	6.289634
Fuel.Type	2.529080
Vehicle.Class	10.026738
Transmission	10.102218

Engine Size Vs Cylinders

04

Multilinear Regression Models

Akaike Information Criterion (AIC)

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Fuel.Consumption.Comb..L.100.km. +
    Fuel.Type + Transmission + Vehicle.Class + Fuel.Consumption.Hwy..L.100.km. +
    Fuel.Consumption.City..L.100.km. + Cylinders + Engine.Size.L.,
    data = noOutliersDF2)
```

Backward Elimination

Forward Selection

Stepwise Selection

```
Call:
lm(formula = CO2.Emissions.g.km. ~ Fuel.Consumption.Comb..L.100.km. +
    Fuel.Type + Transmission + Vehicle.Class + Fuel.Consumption.Hwy..L.100.km. +
    Fuel.Consumption.City..L.100.km. + Cylinders + Engine.Size.L.,
    data = noOutliersDF2)
```

05

Anova Analysis

Anova

Fit 1: with Engine Size

Analysis of Variance Table

Response: CO2.Emissions.g.km.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Engine.Size.L.	1	17088330	17088330	4.8900e+05	< 2.2e-16	***
Fuel.Consumption.City..L.100.km.	1	5929123	5929123	1.6967e+05	< 2.2e-16	***
D	1	111939	111939	3.2033e+03	< 2.2e-16	***
X	1	12776	12776	3.6560e+02	< 2.2e-16	***
COMPACT	1	10954	10954	3.1347e+02	< 2.2e-16	***
SUV...SMALL	1	6986	6986	1.9991e+02	< 2.2e-16	***
MID.SIZE	1	23969	23969	6.8589e+02	< 2.2e-16	***
TWO.SEATER	1	2608	2608	7.4644e+01	< 2.2e-16	***
MINICOMPACT	1	554	554	1.5849e+01	6.930e-05	***
SUBCOMPACT	1	19139	19139	5.4768e+02	< 2.2e-16	***
FULL.SIZE	1	53292	53292	1.5250e+03	< 2.2e-16	***
STATION.WAGON...SMALL	1	15669	15669	4.4839e+02	< 2.2e-16	***
SUV...STANDARD	1	1222	1222	3.4975e+01	3.499e-09	***
VAN...CARGO	1	694	694	1.9862e+01	8.455e-06	***
VAN...PASSENGER	1	956	956	2.7366e+01	1.733e-07	***
PICKUP.TRUCK...STANDARD	1	5370	5370	1.5366e+02	< 2.2e-16	***
MINIVAN	1	2000	2000	5.7238e+01	4.361e-14	***
SPECIAL.PURPOSE.VEHICLE	1	279	279	7.9761e+00	0.004753	**
STATION.WAGON...MID.SIZE	1	4332	4332	1.2396e+02	< 2.2e-16	***
Residuals	6920	241822	35			

Fit 2: with Cylinder

Analysis of Variance Table

Response: CO2.Emissions.g.km.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Cylinders	1	16296093	16296093	4.6783e+05	< 2.2e-16	***
Fuel.Consumption.City..L.100.km.	1	6722893	6722893	1.9300e+05	< 2.2e-16	***
D	1	114523	114523	3.2877e+03	< 2.2e-16	***
X	1	10721	10721	3.0779e+02	< 2.2e-16	***
COMPACT	1	10508	10508	3.0166e+02	< 2.2e-16	***
SUV...SMALL	1	6282	6282	1.8035e+02	< 2.2e-16	***
MID.SIZE	1	23654	23654	6.7905e+02	< 2.2e-16	***
TWO.SEATER	1	2545	2545	7.3062e+01	< 2.2e-16	***
MINICOMPACT	1	488	488	1.4000e+01	0.0001843	***
SUBCOMPACT	1	19098	19098	5.4826e+02	< 2.2e-16	***
FULL.SIZE	1	53082	53082	1.5239e+03	< 2.2e-16	***
STATION.WAGON...SMALL	1	15809	15809	4.5385e+02	< 2.2e-16	***
SUV...STANDARD	1	1306	1306	3.7483e+01	9.726e-10	***
VAN...CARGO	1	723	723	2.0768e+01	5.274e-06	***
VAN...PASSENGER	1	1200	1200	3.4448e+01	4.580e-09	***
PICKUP.TRUCK...STANDARD	1	4835	4835	1.3879e+02	< 2.2e-16	***
MINIVAN	1	2138	2138	6.1369e+01	5.441e-15	***
SPECIAL.PURPOSE.VEHICLE	1	257	257	7.3907e+00	0.0065725	**
STATION.WAGON...MID.SIZE	1	4813	4813	1.3817e+02	< 2.2e-16	***
Residuals	6920	241048	35			

```
> qf(p=.05, df1=1, df2=6909, lower.tail=FALSE)
[1] 3.842805
> qf(p=.05, df1=1, df2=6920, lower.tail=FALSE)
[1] 3.842803
```

Anova Reduced

Anova Without Vehicle Class

Fit 1 :

```
> abs(qt(0.05/2, 6920, lower.tail=TRUE))
[1] 1.960307
```

```
Analysis of Variance Table

Model 1: CO2.Emissions.g.km. ~ Engine.Size.L. + Fuel.Consumption.City..L.100.km. +
    D + X
Model 2: CO2.Emissions.g.km. ~ Engine.Size.L. + Fuel.Consumption.City..L.100.km. +
    D + X + COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER + MINICOMPACT +
    SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL + SUV...STANDARD +
    VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
    MINIVAN + SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   6935 389846
2   6920 241822 15    148024 282.39 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fit 2:

```
Analysis of Variance Table

Model 1: CO2.Emissions.g.km. ~ Cylinders + Fuel.Consumption.City..L.100.km. +
    D + X
Model 2: CO2.Emissions.g.km. ~ Cylinders + Fuel.Consumption.City..L.100.km. +
    D + X + COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER + MINICOMPACT +
    SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL + SUV...STANDARD +
    VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
    MINIVAN + SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   6935 387785
2   6920 241048 15    146737 280.84 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

06

Conclusion

Final Linear Model

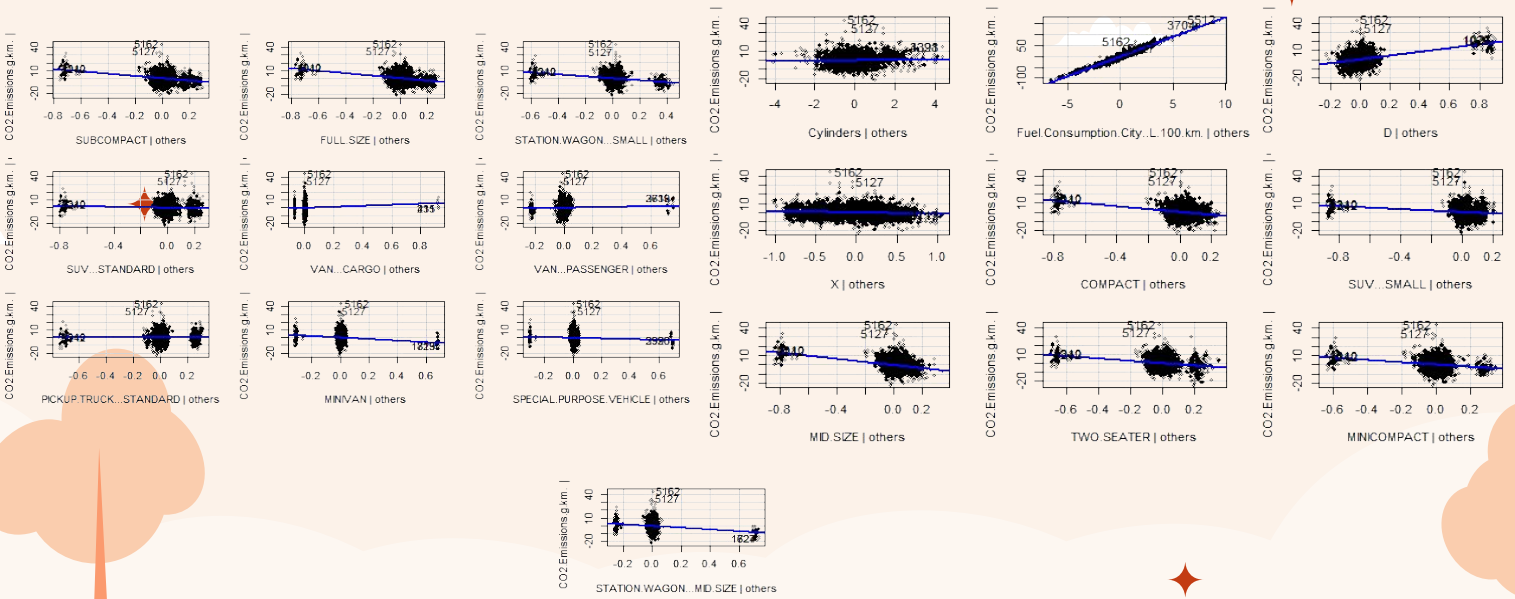
```
Call:
lm(formula = CO2.Emissions.g.km. ~ Cylinders + Fuel.Consumption.City..L.100.km. +
    D + X + COMPACT + SUV...SMALL + MID.SIZE + TWO.SEATER + MINICOMPACT +
    SUBCOMPACT + FULL.SIZE + STATION.WAGON...SMALL + SUV...STANDARD +
    VAN...CARGO + VAN...PASSENGER + PICKUP.TRUCK...STANDARD +
    MINIVAN + SPECIAL.PURPOSE.VEHICLE + STATION.WAGON...MID.SIZE,
    data = nooutliersDF2)

Residuals:
    Min      1Q  Median      3Q     Max
-22.583  -3.926  -0.314   3.544  44.158

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     40.75706    0.70798  57.568  < 2e-16 ***
Cylinders                        0.37907    0.08004   4.736 2.22e-06 ***
Fuel.Consumption.City..L.100.km. 17.88895    0.05062 353.417  < 2e-16 ***
D                               21.09429    0.53653  39.316  < 2e-16 ***
X                               -1.96512    0.18059 -10.881  < 2e-16 ***
COMPACT                         -15.89446    0.54449 -29.192  < 2e-16 ***
SUV...SMALL                      -7.89868    0.51896 -15.220  < 2e-16 ***
MID.SIZE                        -16.25583    0.53818 -30.205  < 2e-16 ***
TWO.SEATER                      -13.38506    0.58945 -22.708  < 2e-16 ***
MINICOMPACT                     -12.94830    0.62155 -20.832  < 2e-16 ***
SUBCOMPACT                      -15.35453    0.56699 -27.081  < 2e-16 ***
FULL.SIZE                       -15.82483    0.56195 -28.160  < 2e-16 ***
STATION.WAGON...SMALL           -12.13207    0.62808 -19.316  < 2e-16 ***
SUV...STANDARD                   -3.52665    0.54636  -6.455 1.16e-10 ***
VAN...CARGO                       6.83230    1.77621   3.847 0.000121 ***
VAN...PASSENGER                   4.07212    1.00935   4.034 5.53e-05 ***
PICKUP.TRUCK...STANDARD           0.38479    0.56858   0.677 0.498589
MINIVAN                          -9.78644    0.86223 -11.350  < 2e-16 ***
SPECIAL.PURPOSE.VEHICLE          -5.23061    0.87420  -5.983 2.30e-09 ***
STATION.WAGON...MID.SIZE        -11.27962    0.95958 -11.755  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.902 on 6920 degrees of freedom
Multiple R-squared:  0.9898,    Adjusted R-squared:  0.9897
F-statistic: 3.519e+04 on 19 and 6920 DF,  p-value: < 2.2e-16
```

Multilinear Regression Plots



07

Future Direction

Future Direction

- Look at the secondary Distribution
- Consider looking at bigger datasets/other datasets with other car brands and more counts of other discrete variables
- Consider impact of Influential points

Thank You