# MA678 Midterm Project

Yujia Wang

December 12, 2021


## Abstract

The purpose of this report is to explore the impact of various variables of fuel vehicles on vehicle emissions CO2. By finding indicators of low CO2 emissions, we can guide car companies to make more environmentally friendly cars. By establishing a multilevel model and analyzing fixed effect and random effect, I found that transmission: AS7, A5, fuel type: Z, X, and especially D are indicators of low CO2 emissions. Manufacturing engines with a smaller engine size under the a constant number of cylinders, choosing a car with a small fuel consumption per unit or with a small vehicle class can also reduce CO2 emissions, which helps companies contribute to our environment and establish a good brand image to attract consumers.


## Introduction

The bigger the company, the more they care about their reputation and take on more social responsibilities. CO2 quantity is an important index to identify whether a car is "environmental friendly" or not. So in this project, I focus on finding what variables affect CO2 emission of cars. Finding greener parts can guide companies in developing cleaner cars. Also, a good reputation can attract consumers. For example, Tesla is using new energy vehicles as their selling point. In conclusion, I will fit multilevel models in order to find out the type of car that consumes less carbon dioxide.

## Method

### Data Cleaning and Combining

My main dataset is about CO2 emissions of cars from Kaggle with different features, such as company, vehicle class, engine size, fuel type, transmission type, fuel consuption and so on. First, in order to perform EDA, I filtered and merged co2 emission of cars data set and Government of Canada's new energy vehicle data set, called co2_combine. Second, in order to perform model fitting, I encode the vehicle classes in the Kaggle data set from their names to their weights according to Wikipedia, so that this column of data can be fitted into the model, called co2_final. Third, for the convenience of more EDA data and observation, I categorized the 27 types of transmission into 5 categories, called co2_final2. The instructions for the above three steps are in the appendix.

### Exploratory data analysis

The left plot of figure 1 shows that the top 5 companies with the most data are BMW, CHEVROLET, FORD, MERCEDES-BENZ, and PORSCHE, so the final results of the model fitting have more reference value for these companies. In addition, the most frequently used transmission of the company is the AS type, followed by the A type and the M type. The right plot of figure 1 shows that companies like to make cars that use fuel X and fuel Z.

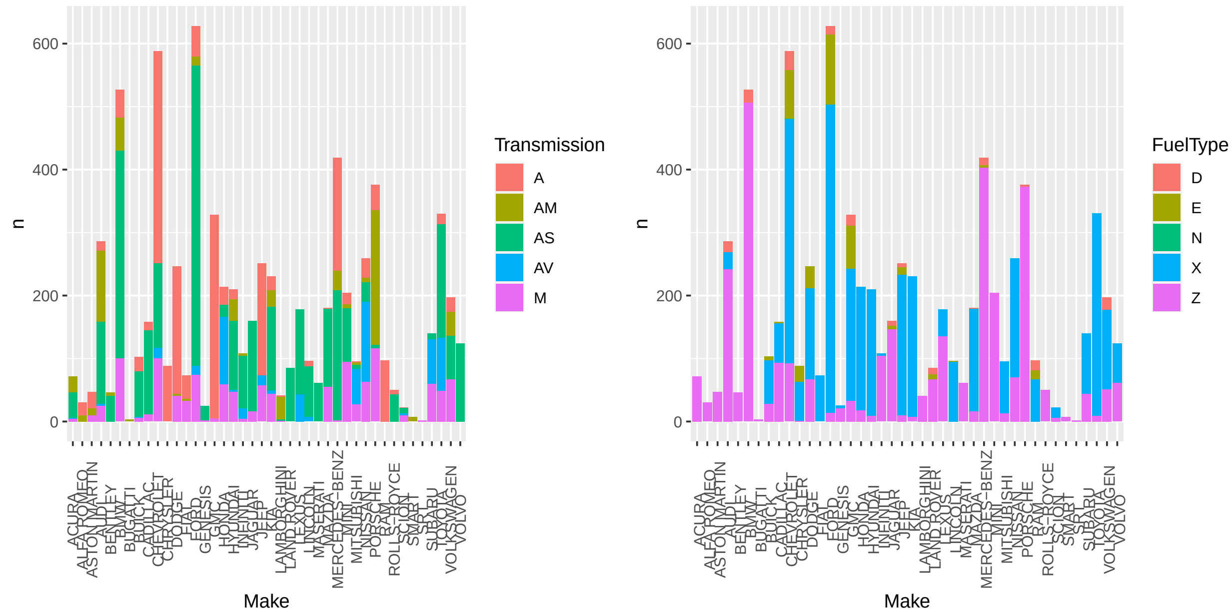



Figure 1: Vehicle brands of different transmissons and fuel types

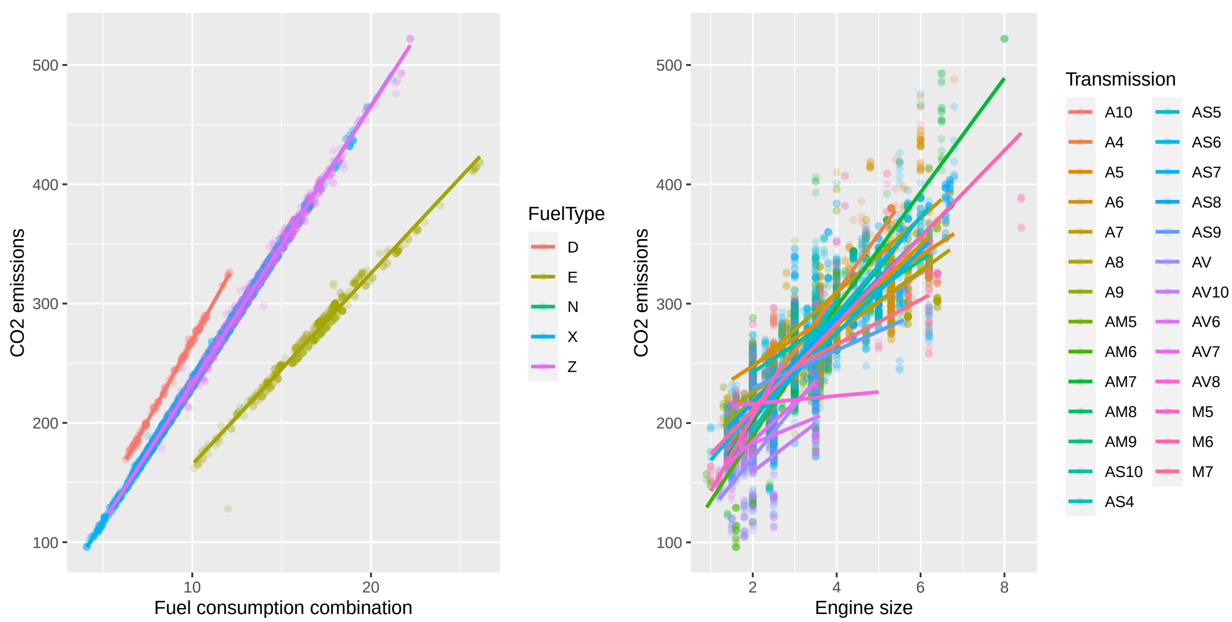



Figure 2: Determining whether use varying slope and intercept

### Model Fitting

In the fixed effect part, first, the size of VehicleClass_Value is related to CO2. Because the model of a car is closely related to its weight, the CO2 emissions released by different weights are different. Second, the size of the engine affects the driving performance of the car, making the release of co2 different. There is a variable called Cylinder in the data set, but we know that $V_{engine} = V_{cylinders} \cdot N_{cylinders}$. So we only choose EngineSize, because it is highly related to Cylinders. Third, fuel consumption must be related to the release of co2. Here I use FC_Comb, which is sensible because it's defined as 55% of the fuel consumption on urban roads combined with 45% of the fuel consumption on highways.

In the random effect part, first, the engine power is transmitted to the driving wheel through transmission. So engine size can be represented by transmission classification. Second, due to the different types of fuel, the fuel consumed by the car is different under other conditions. So FC_Comb can be represented by fuel type classification. According to the variables have different slopes and intercepts in figure 2 from the EDA part above, I choose varying slope and varying intercept.

Finally, I compare the CO2Emissions density plots of FuelType or Transmission with log(CO2Emissions) density plots. The plots are in the More EDA part of Appendix. It can be seen that the original data has only a slight right deviation, and after the log, the data has not improved. So I chose to use the original CO2Emission.

```
model1 <- lmer(CO2Emissions ~ VehicleClass_Value + EngineSize + FC_Comb
               + (1+EngineSize|Transmission) + (1+FC_Comb|FuelType),
               data=co2_final)
```

## Result

### Formula

$$CO2Emissions = 1.54 + 0.00007 \cdot VehicleClassValue + 0.24 \cdot EngineSize + 21.11 \cdot FC_Comb + n_j + i_j + \epsilon$$
$$n_j \sim N(0, \sigma_a^2), i_j \sim N(0, \sigma_b^2)$$

, where $n_j$ and $i_j$ are random effects: $(1 + EngineSize|Transmission)$, $(1 + FC_Comb|FuelType)$

### Fixed effect

| | Estimate | Std. Error | df | t value | Pr(>\|t\|) |
|---|---|---|---|---|---|
| (Intercept) | 1.538e+00 | 8.064e-01 | 2.804e+00 | 1.907 | 0.159011 |
| VehicleClass_Value | 7.075e-05 | 1.975e-05 | 7.139e+03 | 3.582 | 0.000344 *** |
| EngineSize | 2.388e-01 | 1.145e-01 | 2.557e+01 | 2.085 | 0.047207 * |
| FC_Comb | 2.111e+01 | 2.092e+00 | 3.997e+00 | 10.088 | 0.000545 *** |

### Random effect

```
## [[1]]
```

```
##
## [[2]]
```

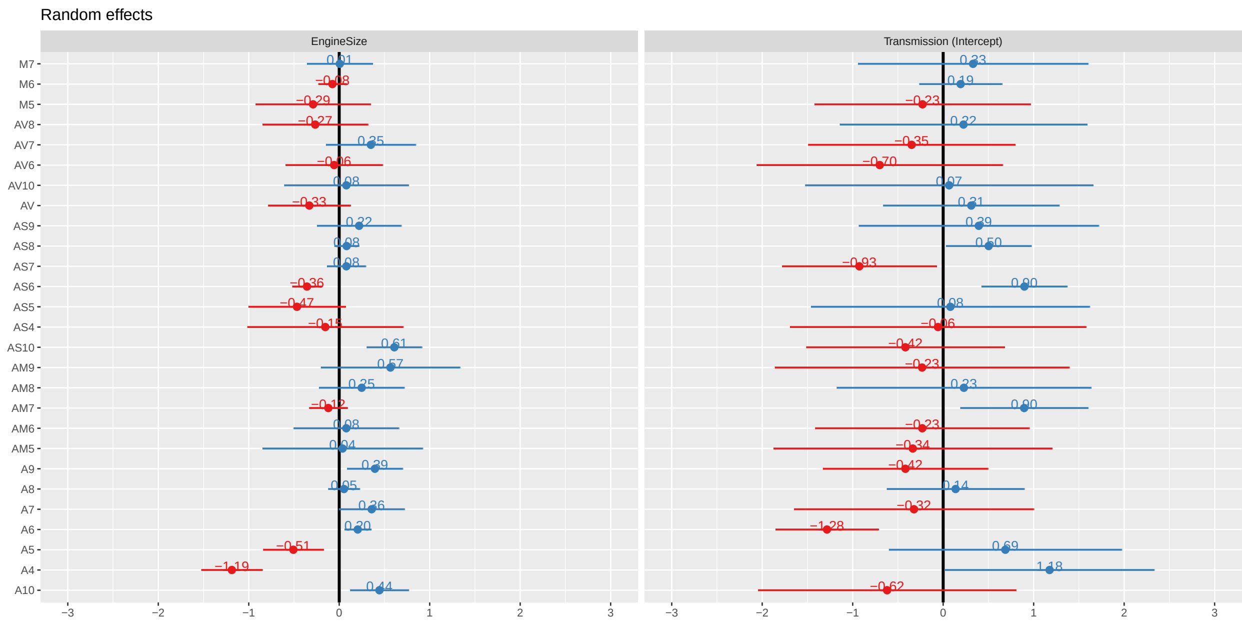



Figure 3: Random effect

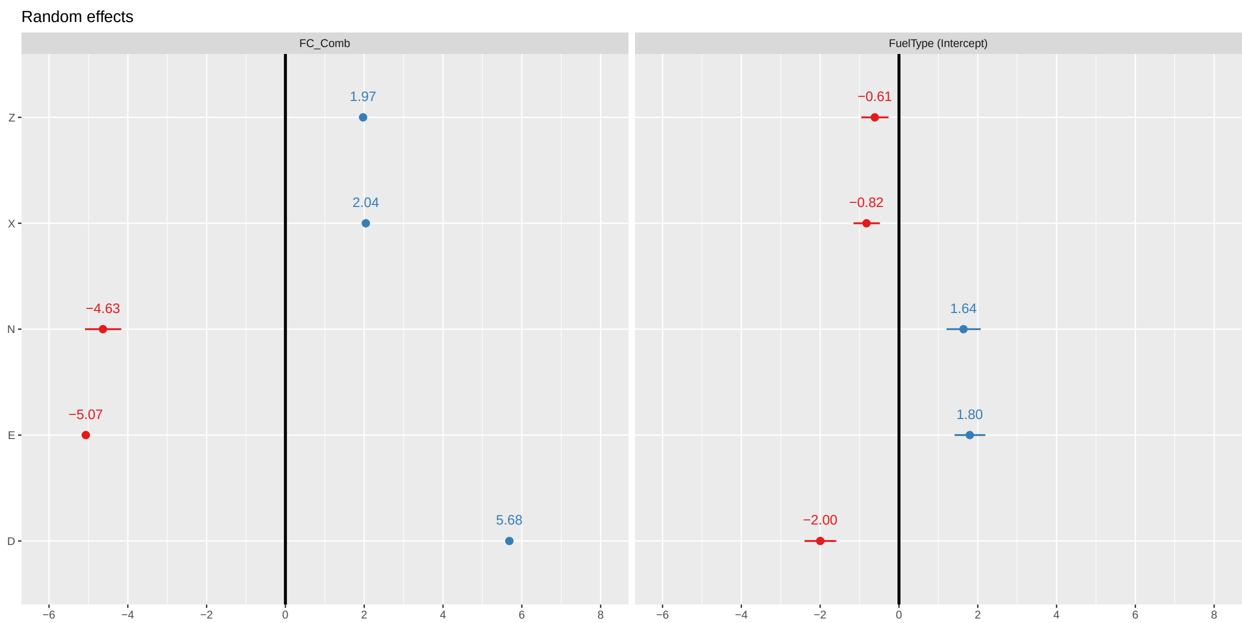



Figure 4: Random effect

## Discussion

According to the fixed effect result, it can be seen that the three variables are all positively correlated with CO2Emissions and there p-values are all significant. For VehicleClass Value, increasing every 1 lb of vehicle weight, co2 will release 0.00007 more (in grams per kilometre). Because this coefficient is too small, it indicates that the model of the vehicle has a small impact on CO2 emissions. This result refutes people's inherent thinking that the larger the vehicle, the more CO2 released. For EngineSize, increasing every 1 Litre of engine, co2 will release 0.24 more (in grams per kilometre). The size of a car engine is usually closely related to the power of the engine. The power is large, the wheel speed increases, the car speed increases, and the release of co2 is also large. For FC_Comb, consuming every 1 L/100 km of fuel more, co2 will release 21.11 more (in grams per kilometre). This is in line with our common sense, because the combustion of fuel produces co2.

According to the random effect result in varying slope, it doesn't show an overall significant difference between various engine size. But in detail, engine size of transmission AS6, AS10, A9, A6, A5, A4, A10 have true effect on slope. In contrast, FC_Comb of all fuel type have significant effect. According to the random effect result in varying intercept, transmission of AS7, AS6, AM7, A6 have true effect and all fuel type have true effect.

In conclusion, based on the goal is to find types with less CO2 emissions, in the transmission angle, select AS7, A5, meanwhile avoid AS8, AM7. In the fuel type angle, select Z, X, especially D, meanwhile avoid N, E. Moreover, according to the previous EDA, these results are more significative for BMW, CHEVROLET, FORD, MERCEDES-BENZ, PORSCHE, TOYOTA and GMC. This may be a rough result. But I believe this can give vehicle companies a reference significance. Choosing the appropriate transmission and fuel type, manufacturing engines with a smaller engine size under the condition of a constant number of cylinders, choosing a car with a small fuel consumption per unit or with a small vehicle class can reduce CO2 emissions and contribute to our environment, which can also establish a good brand image and deserve the support of consumers.

## Limitation

First, the p-value of intercept in the fixed effect part of the model is not significant. Maybe the variables given in the Kaggle data set I selected are still not sufficient, and the model can fit better. Second, the VehicleClass_Value I use is the maximum weight of each vehicle class. It would be more accurate if the weight of each vehicle class of each company and each model could be found, but it would be difficult to complete.

## Reference

CO2 Emission by Vehicles, https://www.kaggle.com/debajyotipodder/co2-emission-by-vehicles?select=CO2+Emissions_Canada.csv

Fuel consumption ratings - Battery-electric vehicles 2012-2021 (2021-10-22), https://open.canada.ca/data/en/dataset/98f1a129-f628-4ce4-b24d-6f16bf24dd64/resource/026e45b4-eb63-451f-b34f-d9308ea3a3d9?inner_span=True

Car classification, https://en.wikipedia.org/wiki/Car_classification

Vehicle size class, https://en.wikipedia.org/wiki/Vehicle_size_class

## Appendix

### Explanation of variables

### Defination of variables

| column names | explanation |
|---|---|
| Make | Company of the vehicle |
| Model | Car Model |
| VehicleClass | Class of vehicle depending on their utility, capacity and weight |
| VehicleClass_Value | The maxinum weight of each vehicle class in lb |
| EngineSize | Size of engine used in Litre |
| Cylinders | Number of cylinders |
| Transmission | Transmission type with number of gears |
| FuelType | Type of Fuel used |
| FC_City | Fuel consumption in city roads (L/100 km) |
| FC_Hwy | Fuel consumption in highways (L/100 km) |
| FC_Comb | The combined fuel consumption (55% city, 45% highway) in L/100 km |
| CO2Emissions | The tailpipe emissions of carbon dioxide (in grams per kilometre) for combined city and highway driving |

### Defination of vehicle class value

| Vehicle Class | Defination |
|---|---|
| (Sedans) | |
| MINICOMPACT | < 85 cubic feet (2,405 l) |
| SUBCOMPACT | 85–99.9 cubic feet (2,405–2,830 l) |
| COMPACT | 100–109.9 cubic feet (2,830–3,110 l) |
| MID-SIZE | 110–119.9 cubic feet (3,115–3,395 l) |
| FULL-SIZE | 120 cubic feet (3,400 l) |
| (Station wagons) | |
| STATION WAGON – SMALL | < 130 cubic feet (3,680 l) |
| STATION WAGON - MID-SIZE | 130–159 cubic feet (3,680–4,500 l) |
| TWO-SEATER | Unknown. It is defined as Roadster or Sports car so I choose 4,500 l. |
| (Trucks) | |
| PICKUP TRUCK – SMALL | <6,000 lb (2,700 kg) |
| PICKUP TRUCK – STANDARD | 6,000–8,500 lb (2,700–3,850 kg) |
| SUV – SMALL | Sport utility vehicles(small) < 6,000 l (2,700 kg) |
| SUV – STANDARD | Sport utility vehicles(standard) 6,000–10,000 l (2,700–4,550 kg) |
| MINIVAN | < 8,500 l (3,850 kg) |
| VAN – CARGO | < 8,500 l (3,850 kg) |
| VAN – PASSENGER | < 10,000 l (4,550 kg) |
| SPECIAL PURPOSE VEHICLE | < 8,500 l (3,850 kg) |

## Defination of transmission type

| Transmission | Defination |
|---|---|
| A | A4, A5, A6, A7, A8, A9, A10 |
| AM | AM5, AM6, AM7, AM8, AM9 |
| AS | AS4, AS5, AS6, AS7, AS8, AS9, AS10 |
| AV | AV, AV6, AV7, AV8, AV10 |
| M | M5, M6, M7 |

## More EDA

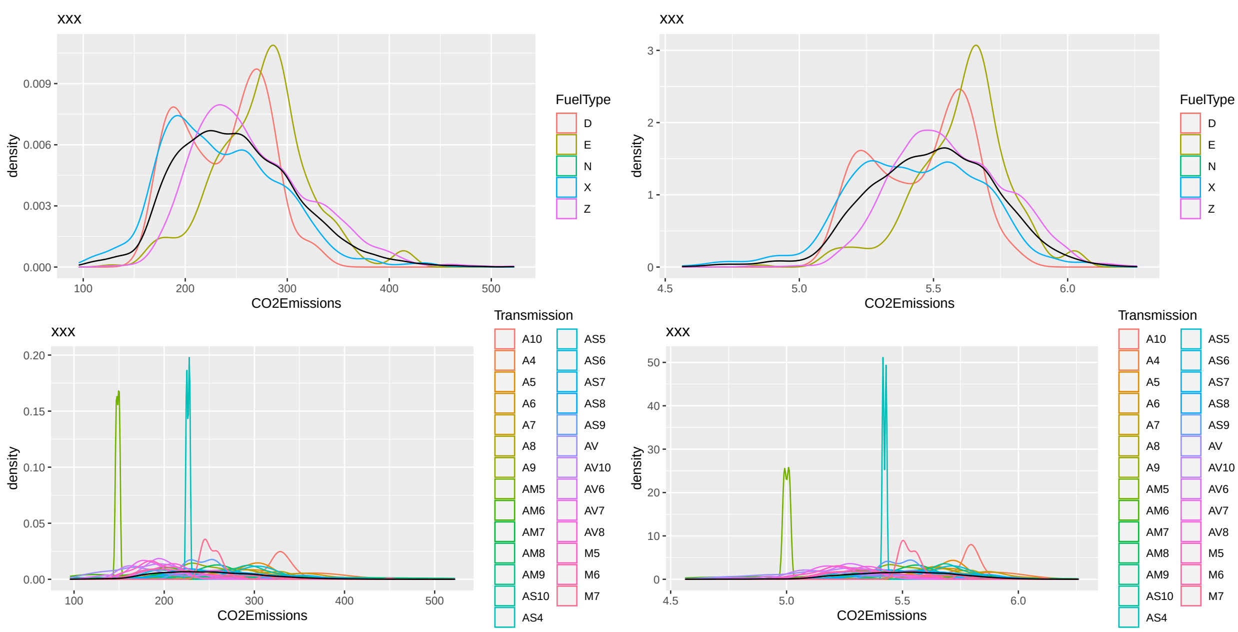



Figure 5: Density of fuel type and transmission using original data and log data

## Model checking

## Residual plot and Q-Q plot

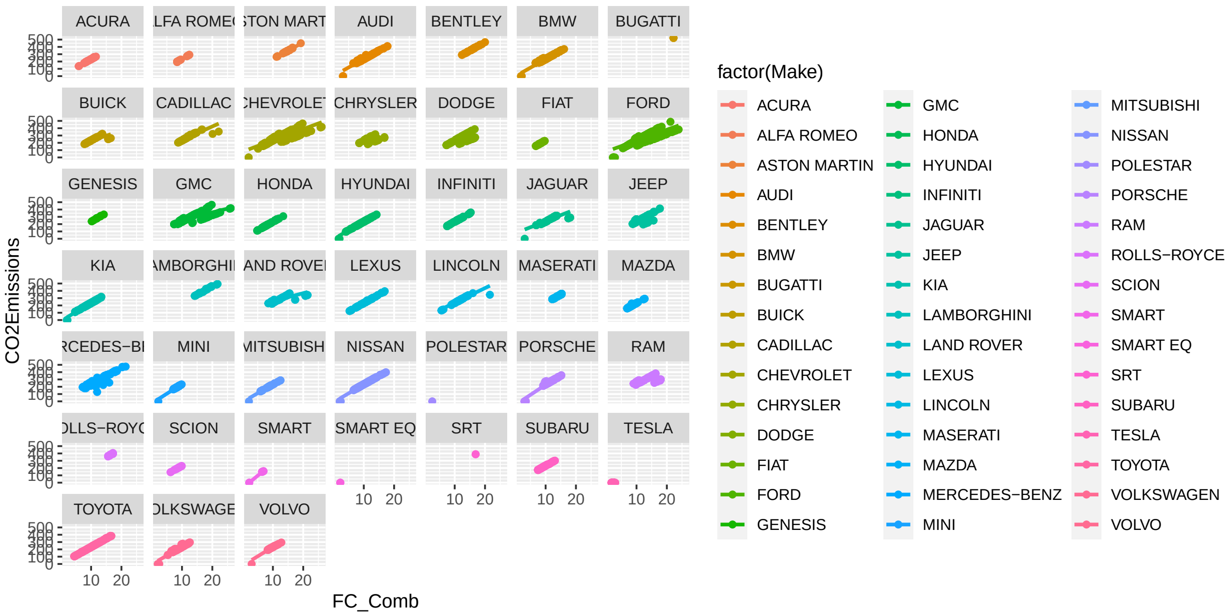



Figure 6: CO2 emission vs. fuel consumption of different companies

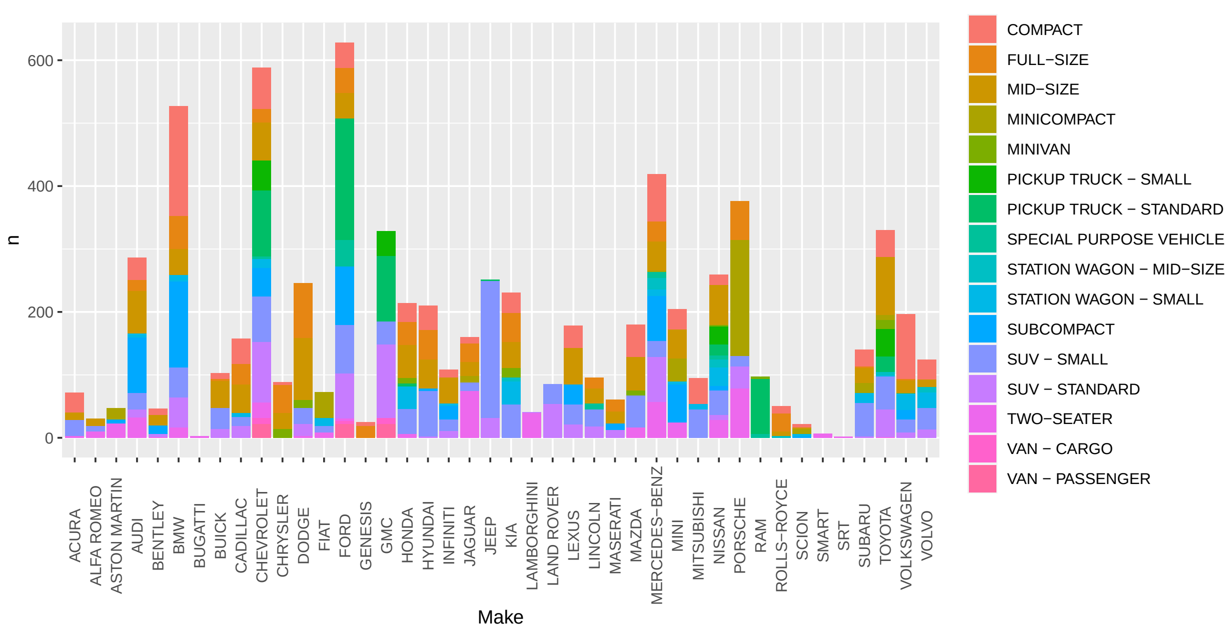



Figure 7: Different companies vs. different vehicle class

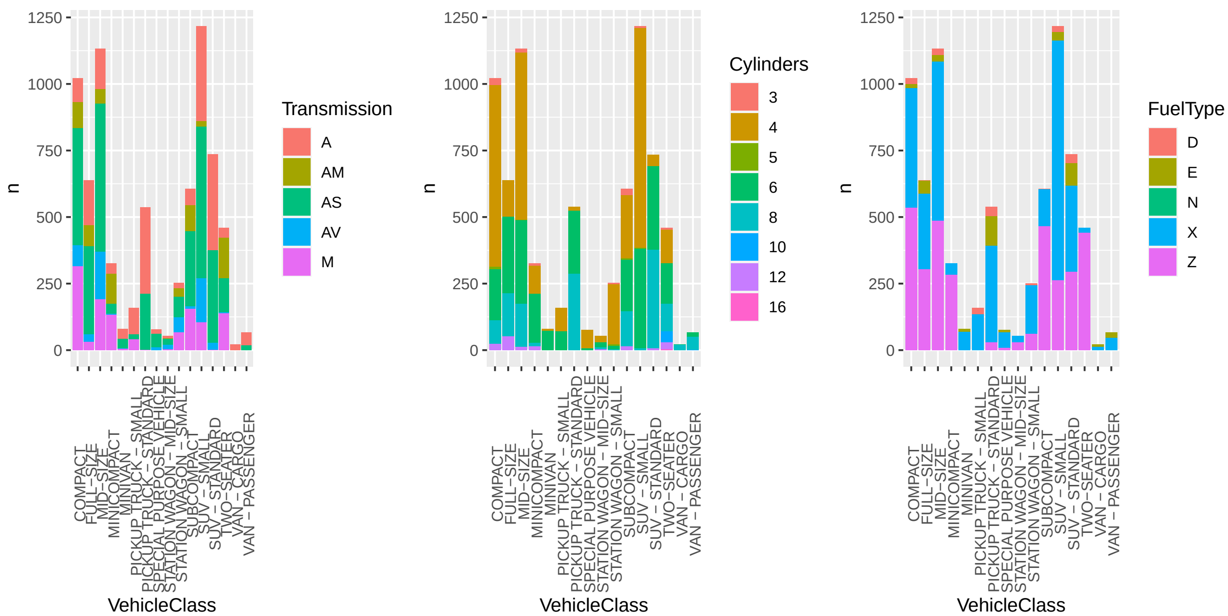
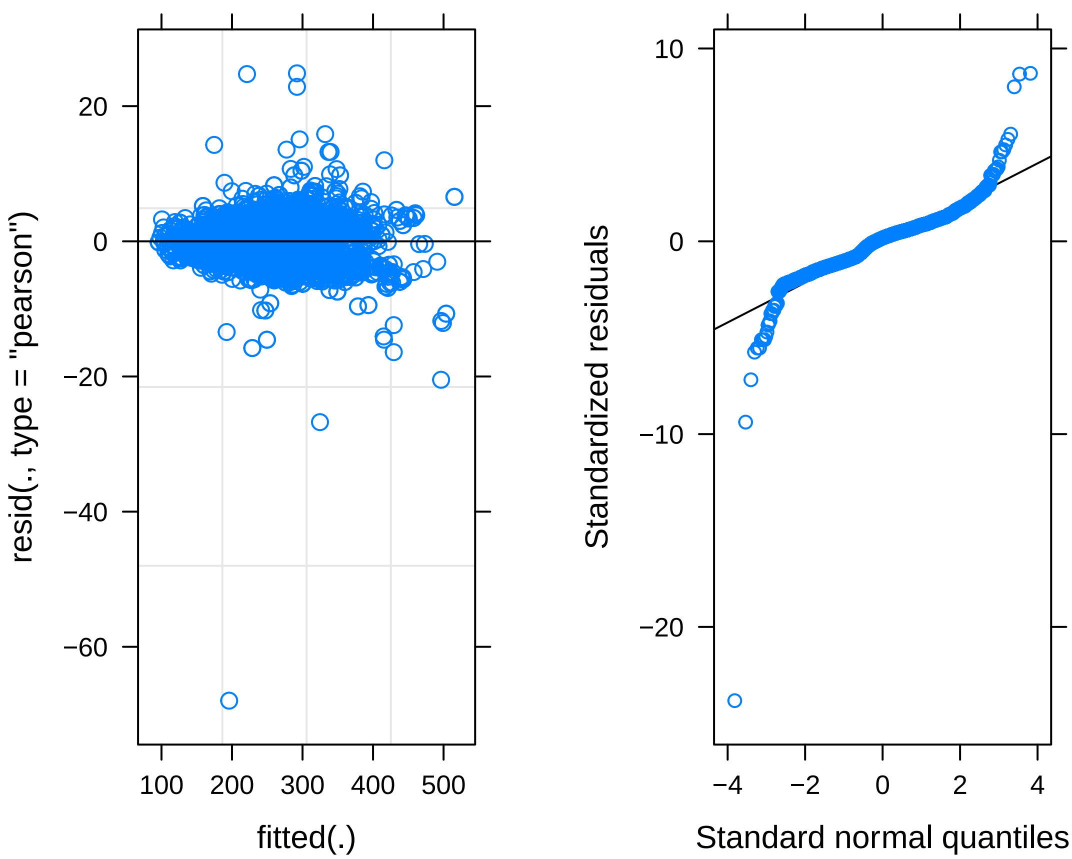



Figure 8: Different vehicle class vs. different transmission, cylinders and fuel type



## Model result

### Random effects result

```
##                   fixef(model1)
## (Intercept)        1.537735e+00
## VehicleClass_Value 7.075298e-05
## EngineSize         2.387549e-01
## FC_Comb            2.110719e+01
```

```
##   (Intercept)   FC_Comb
## D  -1.9977379  5.684262
## E   1.7973598 -5.065594
## N   1.6378368 -4.630876
## X  -0.8241011  2.040919
## Z  -0.6133574  1.971288
```

```
##      (Intercept)  EngineSize
## A10  -0.61966991  0.44422303
## A4    1.17612760 -1.18674551
## A5    0.68721544 -0.50701529
## A6   -1.28420915  0.20461996
## A7   -0.32256847  0.35861634
## A8    0.13727682  0.05256963
## A9   -0.41734339  0.39435703
## AM5  -0.33571300  0.03696100
## AM6  -0.23084568  0.07772976
## AM7   0.89519008 -0.12028956
## AM8   0.22879208  0.24837290
## AM9  -0.23302516  0.56593517
## AS10 -0.41748966  0.60881021
## AS4  -0.05709025 -0.15392152
## AS5   0.07980061 -0.46688463
## AS6   0.89761911 -0.35628158
## AS7  -0.92645526  0.07879733
## AS8   0.50267321  0.08185949
## AS9   0.39364194  0.21986588
## AV    0.31035242 -0.33058747
## AV10  0.06583443  0.07886587
## AV6  -0.70131213 -0.05637178
## AV7  -0.34849596  0.34969630
## AV8   0.22446217 -0.26541136
## M5   -0.22813985 -0.28873305
## M6    0.19313406 -0.07510119
## M7    0.33023790  0.00606305
```

### Full result

```
## MODEL INFO:
## Model type: Linear Mixed Model (LMM)
## = Hierarchical Linear Model (HLM)
## = Multilevel Linear Model (MLM)
##
```

```
## Formula: CO2Emissions ~ VehicleClass_Value + EngineSize + FC_Comb + (1 + EngineSize | Transmission) +
##
## Level-1 Observations: N = 7385
## Level-2 Groups/Clusters: Transmission, 27; FuelType, 5
##
## MODEL FIT:
## AIC = 36634.210
## BIC = 36710.189
## R_(m)² = 0.58627  (Marginal R²: fixed effects)
## R_(c)² = 0.99873  (Conditional R²: fixed + random effects)
## Omega² = 0.99764  (= 1 - proportion of unexplained variance)
##
## ANOVA table:
## -----------------------------------------------------------------------
##                      Sum Sq  Mean Sq  NumDF    DenDF       F      p
## -----------------------------------------------------------------------
## VehicleClass_Value    104.41   104.41   1.00  7139.21   12.83  <.001  ***
## EngineSize             35.38    35.38   1.00    25.57    4.35   .047  *
## FC_Comb               828.18   828.18   1.00     4.00  101.76  <.001  ***
## -----------------------------------------------------------------------
##
## FIXED EFFECTS:
## Outcome variable: CO2Emissions
## ------------------------------------------------------------------------------------
##                           b     S.E.      t      df      p          [95% CI of b]
## ------------------------------------------------------------------------------------
## (Intercept)            1.538  (0.806)   1.91     2.8   .159       [-1.134,  4.209]
## VehicleClass_Value     0.000  (0.000)   3.58  7139.2  <.001  ***  [ 0.000,  0.000]
## EngineSize             0.239  (0.115)   2.08    25.6   .047  *    [ 0.003,  0.474]
## FC_Comb               21.107  (2.092)  10.09     4.0  <.001  ***  [15.296, 26.918]
## ------------------------------------------------------------------------------------
## 'df' is estimated by Satterthwaite approximation.
##
## Standardized coefficients: CO2Emissions
## -----------------------------------------------------------------------------
##                      Beta*  S.E.*     t*   df*     p*       [95% CI of Beta]
## -----------------------------------------------------------------------------
## VehicleClass_Value    0.003  0.001   3.58  7139  <.001  ***    [0.001, 0.005]
## EngineSize            0.006  0.003   2.08    26   .047  *      [0.000, 0.011]
## FC_Comb               1.043  0.103  10.09     4  <.001  ***    [0.756, 1.331]
## -----------------------------------------------------------------------------
##
## RANDOM EFFECTS:
## -------------------------------------------------
##   Cluster       K    Parameter    Variance    ICC
## -------------------------------------------------
##   Transmission   27  (Intercept)   0.69775  0.06008
##                      EngineSize    0.20505
##   FuelType        5  (Intercept)   2.77694  0.23912
##                      FC_Comb      21.87454
##   Residual                         8.13850
## -------------------------------------------------
```