I just found out that [here](https://www.eea.europa.eu/data-and-maps/data/co2-cars-emission-16) the EU made large data sets available about every newly registered car in the EU (currently from 2012 to 2018).

The data set from 2018 has over 15 million rows and many rows correspond to a single registered car. The following analysis uses a smaller data set that aggregated the original huge data sets by grouping on unique combinations of country, year, car model (more precisely the consumer name cn, which still yields 46609 different models), and fuel type. The aggregated data set is available on my Github repository [here](https://github.com/skranz/eu_cars_co2).

OK, let’s first load the data set using the [fst](https://www.fstpackage.org/) package.

library(fst)

dat = fst::read\_fst("cars.fst")

nrow(dat)

## [1] 260639

head(dat)

## year firm ms mp man mh mms cn ft it co2

## 1 2012 andere AT aa-iva aa-iva andere 60849 petrol 270

## 2 2012 andere AT aa-iva aa-iva andere COOLCAR ELEKTRO electric 0

## 3 2012 andere AT aa-iva aa-iva andere SYLVA RIOT petrol 183

## 4 2012 audi AT aa-iva aa-iva audi A8 L petrol 290

## 5 2012 audi AT aa-iva aa-iva audi AUDI A1 QUATTRO petrol 199

## 6 2012 audi AT aa-iva aa-iva audi AUDI Q3 petrol 256

## mass cm3 power elec q wltp red\_nedc red\_wltp

## 1 1795 2351 92 NaN 1 NA NA NA

## 2 1515 NaN 20 NaN 1 NA NA NA

## 3 575 1340 145 NaN 1 NA NA NA

## 4 2130 6299 368 NaN 1 NA NA NA

## 5 1580 1984 188 NaN 1 NA NA NA

## 6 1760 2480 228 NaN 1 NA NA NA

The column q describes the number of newly registered cars. Let us first take a look how the total number has developed over time:

library(ggplot2)

library(dplyr)

dat %>%

group\_by(year) %>%

summarize(Mio\_cars = sum(q,na.rm=TRUE) / 1e6) %>%

ggplot(aes(x=year,y=Mio\_cars)) +

geom\_line() + geom\_point()

We see that in 2018 there were roughly 25% more newly registered cars than in 2012.

The column co2 measures how much grams CO2 per km the car emits using the [New European Driving Cycle (NEDC)](https://en.wikipedia.org/wiki/New_European_Driving_Cycle), which is actually an older procedure that is currently replaced by the [WLTP](https://ec.europa.eu/jrc/en/publication/eur-scientific-and-technical-research-reports/nedc-wltp-effect-type-approval-co2-emissions-light-duty-vehicles).  
Let us take a look at the average co2 emissions per km over time:

library(ggplot2)

library(dplyr)

dat %>%

group\_by(year) %>%

summarize(co2 = weighted.mean(co2,q,na.rm=TRUE)) %>%

ggplot(aes(x=year,y=co2)) +

geom\_line() + geom\_point()

While on average new cars got considerably more CO2 efficient from 2012 to 2015, the trend did not continue. Rather CO2 performance slightly deteriorated from 2015 to 2018. One factor could be [Dieselgate](https://en.wikipedia.org/wiki/Volkswagen_emissions_scandal) brought to light in 2015. Let us look at the fuel type composition of new cars:

# We have 15 different categories of fuel types

unique(dat$ft)

## [1] "petrol" "electric" "diesel"

## [4] "e85" "lpg" "ng-biomethane"

## [7] "cng" "petrol-electric" "diesel-electric"

## [10] "unknown" "biodiesel" "hydrogen"

## [13] "other" "petrol-gas" "ng"

# Let us lump all but the 2 most common together

# using fct\_lump from the forcats packae

library(forcats)

dat = dat %>%

mutate(fuel2 = fct\_lump(ft, 2))

dat\_fuel = dat %>%

group\_by(year) %>%

mutate(q.year = sum(q, na.rm=TRUE)) %>%

group\_by(year, fuel2) %>%

summarize(

mass = weighted.mean(mass, q, na.rm=TRUE),

co2 = weighted.mean(co2, q, na.rm=TRUE),

share = sum(q)/first(q.year)

)

ggplot(dat\_fuel, aes(x=year,y=share)) +

geom\_line() + geom\_point() + facet\_wrap(~fuel2)

We see how the share of diesel cars has been declining (the dent in 2013 is probably due to missing fuel type classification of some cars in that year which appear under Other) while the share of petrol cars has steadily increased. Other fuel types play with a total of 4.2% in 2018 no substantial role yet.

Let us take a look at the average CO2 emission per km for diesel and petrol.

dat\_fuel = dat\_fuel %>%

filter(fuel2!="Other")

ggplot(dat\_fuel, aes(x=year,y=co2, color=fuel2)) +

geom\_line() + geom\_point()

Diesel cars have in every year on average always lower CO2 emissions than petrol cars. The differences are quite small in most years though. However, the average car sold with a diesel engine may differ in several dimensions from one with petrol engine.

Let us compare the average mass (in kg) of the cars:

ggplot(dat\_fuel,aes(x=year,y=mass, color=fuel2)) +

geom\_line() + geom\_point()

The average car with diesel engine is roughly 300 kg heavier than one with petrol engine. Part of the extra mass is due to the fact that diesel engines are just heavier than petrol engines. Yet, that is unlikely to account for all 300 kg. Larger, more heavy car models seem to sell more often with diesel engines than smaller cars. To move more mass, more fuel must be burned. Holding the car model fixed (but not the fuel type), we would thus expect to find a larger CO2 reduction when switching from petrol to diesel than suggested in the plot above.

To explore this in detail, we run the following 3 regressions

dat\_reg = dat %>%

filter(ft %in% c("diesel","petrol")) %>%

filter(year == 2018) %>%

mutate(model = paste0(firm," ", cn))

reg1 = lm(co2 ~ ft, weights = q,data = dat\_reg)

reg2 = lm(co2 ~ ft+mass, weights=q, data = dat\_reg)

# for regression with model fixed effects

# use felm instead of lm for speed and memory reasons

library(lfe)

reg3 = felm(co2 ~ ft | model,weights = dat\_reg$q, data=dat\_reg)

library(stargazer)

stargazer(reg1, reg2, reg3,type = "html")

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  | *Dependent variable:* | | |
|  |  | | |
|  | co2 | | |
|  | *OLS* | | *felm* |
|  | (1) | (2) | (3) |
|  | | | |
| ftpetrol | 1.856\*\*\* | 24.341\*\*\* | 17.213\*\*\* |
|  | (0.238) | (0.169) | (0.090) |
|  |  |  |  |
| mass |  | 0.079\*\*\* |  |
|  |  | (0.0003) |  |
|  |  |  |  |
| Constant | 121.872\*\*\* | -1.231\*\* |  |
|  | (0.187) | (0.479) |  |
|  |  |  |  |
|  | | | |
| Observations | 42,476 | 42,450 | 42,476 |
| R2 | 0.001 | 0.623 | 0.954 |
| Adjusted R2 | 0.001 | 0.623 | 0.928 |
| Residual Std. Error | 440.907 (df = 42474) | 270.844 (df = 42447) | 118.155 (df = 27480) |
| F Statistic | 60.959\*\*\* (df = 1; 42474) | 35,133.410\*\*\* (df = 2; 42447) |  |
|  | | | |
| *Note:* | \*p<0.1; \*\*p<0.05; \*\*\*p<0.01 | | |

The first regression without any control variables finds as in our plot that the average petrol car emits in 2018 1.8g CO2/km more than a diesel car. If we control for the car mass (2nd regression) diesel has a much larger benefit of 24g CO2/km. However, this overestimates the benefits from a diesel engine because it does not account for the fact that a diesel engine causes a higher car mass. The third regression controls for car model specific fixed effects, i.e. we essentially measure the average co2 difference between diesel and petrol engines for the same car models. The 17g CO2/km is almost 10 times the size of the 1.8g CO2/km difference shown in the plot.

The last regression could be misleading if for most cars the model label already differentiates between the fuel type (e.g. a Golf TSI has a petrol engine and a Golf TDI a diesel engine.) The code below verifies however, that over 2/3 of cars sold are from models that have entries for both petrol and diesel engines.

# How many consumer names do we have with diesel and petrol?

dat\_reg %>%

group\_by(model) %>%

mutate(

has\_diesel = "diesel" %in% ft,

has\_petrol = "petrol" %in% ft,

has\_both = has\_diesel & has\_petrol

) %>%

group\_by(has\_both) %>%

summarize(q = sum(q, na.rm=TRUE), count=n())

## # A tibble: 2 x 3

## has\_both q count

##

## 1 FALSE 4312320 26075

## 2 TRUE 10311785 16712

From 2012 to 2018 the share of petrol cars has gone up by roughly 17 percentage points. If in 2018 an extra 17% of all cars would drive with diesel instead of petrol, the calculations above suggest as rough estimate that in 2018 the average CO2 emission would have been lower by 0.17\*17.21 = 2.9 g CO2/km which is roughly only 2.4% of the average 2018 emissions.

Other reasons for the stagnation of CO2 efficiency from 2015 onward could be shifts in consumer preferences towards more heavy or more powerful cars (like SUV). Let us make the simplifying assumption that CO2 efficiency only depends on the car’s mass, engine power, fuel type and a year fixed effect that measures the state of technology. The following code then compares the actual development of CO2 efficiency with an hypothetical pure technological effect that we would get if the average car mass, engine power and fuel type would stay in every year the same as in 2012.

reg\_dat = dat %>%

# filter(ft %in% c("diesel","petrol")) %>%

mutate(Year = as.factor(year))

# Run regression and store year dummies

reg = lm(co2 ~ 0+Year+mass+power+ft ,data=reg\_dat, weights=reg\_dat$q)

year.dummies = coef(reg)[1:7]

plot\_dat = dat %>%

group\_by(year) %>%

summarize(real\_co2 = weighted.mean(co2,q,na.rm=TRUE)) %>%

mutate(technological = real\_co2[1]-year.dummies[1] + year.dummies ) %>%

tidyr::pivot\_longer(c("real\_co2", "technological"),

names\_to = "type", values\_to = "co2")

tail(plot\_dat,4)

## # A tibble: 4 x 3

## year type co2

##

## 1 2017 real\_co2 119.

## 2 2017 technological 117.

## 3 2018 real\_co2 121.

## 4 2018 technological 118.

ggplot(plot\_dat, aes(x=year,y=co2, color=type)) +

geom\_line() + geom\_point() +

# CO2 target before 2020

geom\_hline(yintercept = 130, linetype="dashed") +

# CO2 target in 2020

geom\_hline(yintercept = 95, linetype="dashed")

While the pure technological effect has in total a slightly stronger CO2 reduction than the real reduction (ca. 3 g CO2/km difference) that is a rather small effect. Indeed, it is of a similar size as the previously estimated effect from the switch from diesel to petrol engines.

I added to the plot the overall EU emission targets of 130 g CO2/km before 2020 and 95 g CO2/km in 2020 as a reference (In a second blog post, we will look in much more detail at the EU regulation and what the new target means for particular car companies). We see that even the pure technological measure is in 2018 far away from the new CO2 target.

Overall, being a layman concerning the automobile industry and not knowing anything about engineering, I cannot infer the reason for the slowdown of CO2 reductions. Perhaps after Dieselgate the tests became more accurate and thus tougher. Perhaps car companies did not yet put much effort in reducing CO2 emissions in 2018 because the current target is well reached. Perhaps traditional diesel and petrol technology is reaching its efficiency frontier and only very gradual progress is possible.

Anyway, it seems a fair bet that without a stronger reliance on alternative fuel technologies, it will be very hard to reach the new CO2 target. Let us plot the average CO2 emissions by fuel type for newly registered cars in 2018:

dat %>%

filter(year == 2018) %>%

group\_by(ft) %>%

summarize(

co2 = weighted.mean(co2, q,na.rm = TRUE)

) %>%

ggplot(aes(x=fct\_reorder(ft,-co2), y=co2)) +

geom\_col() + xlab("Fuel Type") +

geom\_hline(yintercept = 95) + coord\_flip()

While natural gas (ng) performs a bit better than diesel or petrol, the average is still above the new emission target. Liquified petrol gas (lpg) performs even a bit worse than gasoline. (Note however that these raw averages do not control for other car characteristics.)

Electric and hydrogen cars are counted as zero emission cars and also hybrid cars (petrol-electric and diesel-electric) are substantially below the new target. It is, of course, quite debatable to classify electric cars for regulatory purposes as zero emission vehicles since much electricity is still generated by burning dirty coal and somewhat less dirty natural gas. (See my [earlier post](https://skranz.github.io/r/2018/11/21/Electric_Cars.html) for a link to an interactive study that compares the total environmental impacts of driving electric and gasoline cars using US data.)

While there is hope that electricity production will become more green, progress is slow. In particular, Germany currently faces the huge problem that construction of new onshore wind power plants has severely dropped in 2019 (see the data [here](https://energy-charts.de/power_inst_de.htm?year=all&period=annual&type=inc_dec)), mainly due to lengthy legal actions against new plants and lack of suitable locations.

Anyway the regulation is as it is. Let us have a look at the development of the shares of those 4 car types:

dat %>%

group\_by(year) %>%

mutate(q.year = sum(q, na.rm=TRUE)) %>%

group\_by(year, ft) %>%

summarize(share = 100\*sum(q, na.rm=TRUE) / first(q.year)) %>%

filter(ft %in% c("hydrogen","electric",

"petrol-electric","diesel-electric")) %>%

ggplot(aes(x=year, y=share)) +

geom\_line() + geom\_point() + facet\_wrap(~ft)

Of these 4 technologies, only electric and petrol-electric hybrid cars seem to be currently relevant in Europe. OK, let’s finish for now. In the next blog post on this topic, I want to use the data set to investigate in more detail the new regulations and the impacts on different producers and car types.