Interactive EDA is nice but customized interactive EDA is even nicer. To celebrate the new CRAN version of my ‘ExPanDaR’ package I prepare a customized variant of ‘ExPanD’ to explore the [U.S. EPA data on fuel economy](http://www.fueleconomy.gov/). Our objective is to develop an interactive display that guides the reader on how to explore the fuel economy data in an intuitive way.

First, let’s load the packages and the data from EPA’s web page. In addition, I prepared a small data set containing the countries of domicile for the car producers with more than 100 different cars listed in the EPA data (covering the vast majority of the EPA sample).

library(tidyverse)

library(ExPanDaR)

# The following two chuncks borrow

# from the raw data code of the

# fueleconomy package by Hadley Wickham,

# See: https://github.com/hadley/fueleconomy

if(!file.exists("vehicles.csv")) {

tmp <- tempfile(fileext = ".zip")

download.file("http://www.fueleconomy.gov/feg/epadata/vehicles.csv.zip",

tmp, quiet = TRUE)

unzip(tmp, exdir = ".")

}

raw <- read.csv("vehicles.csv", stringsAsFactors = FALSE)

countries <- read.csv("https://joachim-gassen.github.io/data/countries.csv",

stringsAsFactors = FALSE)

Next, I merge the data and re-code some factors to be more stable across time as the EDA has made several coding changes over time.

vehicles <- raw %>%

mutate(car = paste(make, model, trany),

mpg\_hwy = ifelse(highway08U > 0, highway08U, highway08),

mpg\_city = ifelse(city08U > 0, city08U, city08)) %>%

left\_join(countries) %>%

select(car, make, country, trans = trany,

year,

class = VClass, drive = drive, fuel = fuelType,

cyl = cylinders, displ = displ,

mpg\_hwy, mpg\_city) %>%

filter(drive != "",

year > 1985,

year < 2020) %>%

mutate(fuel = case\_when(

fuel == "CNG" ~ "gas",

fuel == "Gasoline or natural gas" ~ "hybrid\_gas",

fuel == "Gasoline or propane" ~ "hybrid\_gas",

fuel == "Premium and Electricity" ~ "hybrid\_electro",

fuel == "Premium Gas or Electricity" ~ "hybrid\_electro",

fuel == "Premium Gas and Electricity" ~ "hybrid\_electro",

fuel == "Regular Gas or Electricity" ~ "hybrid\_electro",

fuel == "Electricity" ~ "electro",

fuel == "Diesel" ~ "diesel",

TRUE ~ "gasoline"

),

class = case\_when(

grepl("Midsize", class) ~ "Normal, mid-size",

grepl("Compact", class) ~ "Normal, compact",

grepl("Small Station Wagons", class) ~ "Normal, compact",

grepl("Large Cars", class) ~ "Normal, large",

grepl("Minicompact", class) ~ "Normal, sub-compact",

grepl("Subcompact", class) ~ "Normal, sub-compact",

grepl("Two Seaters", class) ~ "Two Seaters",

grepl("Pickup Trucks", class) ~ "Pickups",

grepl("Sport Utility Vehicle", class) ~ "SUVs",

grepl("Special Purpose Vehicle", class) ~ "SUVs",

grepl("Minivan", class) ~ "(Mini)vans",

grepl("Vans", class) ~ "(Mini)vans"

),

drive = case\_when(

grepl("4-Wheel", drive) ~ "4-Wheel Drive",

grepl("4-Wheel", drive) ~ "4-Wheel Drive",

grepl("All-Wheel", drive) ~ "4-Wheel Drive",

grepl("Front-Wheel", drive) ~ "Front-Wheel Drive",

grepl("Rear-Wheel", drive) ~ "Rear-Wheel Drive"

),

trans = case\_when(

grepl("Automatic", trans) ~ "Automatic",

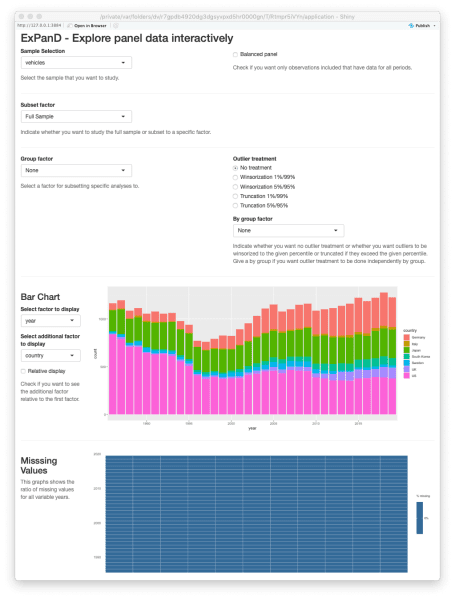
grepl("Manual", trans) ~ "Manual"

)) %>%

na.omit()

Now, the EPA fuel economy data is ready for exploration. Nothing stops you from firing up ‘ExPanD’ right away.

ExPanD(vehicles, cs\_id = "cars", ts\_id = "year")



ExPanD with fuel economy data, standard look

This works but the display is not optimal for what we are after. It includes components that are not needed (e.g., the missing values display as the data do not contain any missing observations) and lacks documentation. We can do better by using some of the new customization features of the ‘ExPanD’ app. See below for a ‘customized’ version that guides the reader a little bit better through the analysis by omitting unneeded displays, adding variable definitions, setting informative defaults and adding explanatory texts between displays.

df\_def <- data.frame(

var\_name = names(vehicles),

var\_def = c("Make, model and transition type indentifying a unique car in the data",

"Make of car",

"Country where car producing firm is loacted",

"Transition type (automatic or manual)",

"Year of data",

"Classification type of car (simplified from orginal data)",

"Drive type of car (Front Wheel, Rear Wheel or 4 Wheel)",

"Fuel type (simplified from orginal data)",

"Number of engine cylinders",

"Engine displacement in liters",

"Highway miles per gallon (MPG). For electric and CNG vehicles this number is MPGe (gasoline equivalent miles per gallon).",

"City miles per gallon (MPG). For electric and CNG vehicles this number is MPGe (gasoline equivalent miles per gallon)."),

type = c("cs\_id", rep("factor", 3), "ts\_id", rep("factor", 3), rep("numeric", 4))

)

html\_blocks <- c(

paste("

",

"By default, this display uses all data from car makes with more",

"than 100 cars in the 'fueleconomy.gov' database.",

"Above, you can limit the analysis to cars from a certain make,",

"class, country, fuel type or other factor present in the data.",

"

"),

paste("

",

"In the display above, remove the check mark to see the absolute",

"number of cars included in the data each year.",

"Also, change the additional factor to see how the distribution",

"of cars across countries, transition types, etc. changes over time",

"

"),

paste("

",

"In the two tables above, you can assess the distributions of the",

"four numerical variables of the data set. Which car has the",

"largest engine of all times?",

"

"),

paste("

",

"Explore the numerical variables across factors. You will see,",

"not surprisingly, that fuel economy varies by car class.",

"Does it also vary by drive type?",

"

"),

paste("

",

"The above two panels contain good news. Fuel economy has",

"increased over the last ten years. See for yourself:",

"Has the size of engines changed as well?",

"

"),

paste("

",

"The scatter plot documents a clear link between engine size",

"and fuel economy in term of miles per gallon.",

"Below, you can start testing for associations.",

"

"),

paste("

",

"Probably, you will want to test for some associations that",

"require you to construct new variables. No problem. Just enter the",

"variable definitions above. Some ideas on what to do:",

"

* Define country dummies (e.g., country == 'US') to see",
* "whether cars from certain countries are less fuel efficient than others.

",

"

* Define a dummy for 4-Wheel drive cars to assess the penalty",
* "of 4-Wheel drives on fuel economy.

",

"

* If you are from a metric country, maybe your are mildly annoyed",
* "by the uncommon way to assess fuel economy via miles per gallon.",
* "Fix this by defining a liter by 100 km measure",
* "(hint: 'l100km\_hwy := 235.215/mpg\_hwy').

",

"

"),

paste("

",

"Above, you can play around with certain regression parameters.",

"See how robust coefficients are across car classes by estimating",

"the models by car class ('subset' option).",

"Try a by year regression to assess the development of fuel economy",

"over time.   
   
",

"If you like your analysis, you can download the configuration",

"and reload it at a later stage using the buttons below.",

"

")

)

cl <- list(

ext\_obs\_period\_by = "2019",

bgbg\_var = "mpg\_hwy",

bgvg\_var = "mpg\_hwy",

scatter\_loess = FALSE,

delvars = NULL,

scatter\_size = "cyl",

bar\_chart\_relative = TRUE,

reg\_x = c("cyl", "displ", "trans"),

scatter\_x = "displ",

reg\_y = "mpg\_hwy",

scatter\_y = "mpg\_hwy",

bgvg\_byvar = "class",

quantile\_trend\_graph\_var = "mpg\_hwy",

bgbg\_byvar = "country",

scatter\_color = "country", bar\_chart\_var2 = "class",

ext\_obs\_var = "mpg\_hwy",

trend\_graph\_var1 = "mpg\_hwy",

trend\_graph\_var2 = "mpg\_city",

sample = "vehicles"

)

ExPanD(vehicles, df\_def = df\_def, config\_list = cl,

title = "Explore the Fuel Economy of Cars in the U.S. Market",

abstract = paste("This interactive display features the [",](https://www.fueleconomy.gov/" \t "_blank)

["fuel economy data provided by the U.S. Environmental Protection Agency](https://www.fueleconomy.gov/" \t "_blank)",

"and allows you to explore the fuel economy of cars in the U.S. market",

"across time and other dimensions. Scroll down and enjoy!"),

components = c(subset\_factor = TRUE,

html\_block = TRUE,

bar\_chart = TRUE,

html\_block = TRUE,

descriptive\_table = TRUE,

ext\_obs = TRUE,

html\_block = TRUE,

by\_group\_bar\_graph = TRUE,

by\_group\_violin\_graph = TRUE,

html\_block = TRUE,

trend\_graph = TRUE,

quantile\_trend\_graph = TRUE,

html\_block = TRUE,

scatter\_plot = TRUE,

html\_block = TRUE,

udvars = TRUE,

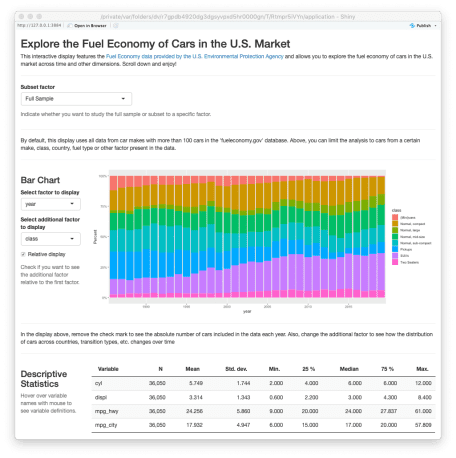
html\_block = TRUE,

regression = TRUE,

html\_block = TRUE),

html\_blocks = html\_blocks

)



ExPanD with fuel economy data, customized look

You can [assess the customized EDA online here](https://jgassen.shinyapps.io/expand_fuel_economy). While customizing ‘ExPanD’ requires extra work, it helps to make the EDA flow better. Fur further information on how to build customized EDAs using ‘ExPanD’ step-by-step, please [refer to the new vignette of the package](https://joachim-gassen.github.io/ExPanDaR/articles/customize_ExPanD.html).