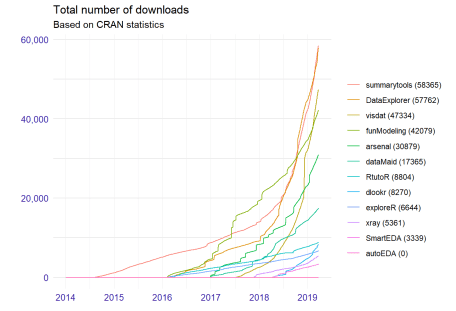
Do you spend a lot of time on data exploration? If yes, then you will like today’s post about AutoEDA

If you ever dreamt of automating the first, laborious part of data analysis when you get to know the variables, print descriptive statistics, draw a lot of histograms and scatter plots – you weren’t the only one. Turns out that a lot of R developers and users thought of the same thing. There are over a dozen R packages for automated Exploratory Data Analysis and the interest in them is growing quickly. Let’s just look at this plot of number of downloads from the official CRAN repository.



Replicate this plot with

stats <- archivist::aread("mstaniak/autoEDA-resources/autoEDA-paper/52ec")

stat <- stats %>%

filter(date > "2014-01-01" ) %>%

arrange(date) %>%

group\_by(package) %>%

mutate(cums = cumsum(count),

packages = paste0(package, " (",max(cums),")"))

stat$packages <- reorder(stat$packages, stat$cums, function(x)-max(x))

ggplot(stat, aes(date, cums, color = packages)) +

geom\_step() +

scale\_x\_date(name = "", breaks = as.Date(c("2014-01-01", "2015-01-01",

"2016-01-01", "2017-01-01",

"2018-01-01", "2019-01-01")),

labels = c(2014:2019)) +

scale\_y\_continuous(name = "", labels = comma) +

DALEX::theme\_drwhy() +

theme(legend.position = "right", legend.direction = "vertical") +

scale\_color\_discrete(name="") +

ggtitle("Total number of downloads", "Based on CRAN statistics")

New tools arrive each year with a variety of functionalities: creating summary tables, initial visualization of a dataset, finding invalid values, univariate exploration (descriptive and visual) and searching for bivariate relationships.

We compiled a list of R packages dedicated to automated EDA, where we describe twelve packages: their capabilities, their strong aspects and possible extensions.

Spoiler alert: currently, automated means simply fast. The packages that we describe can perform typical data analysis tasks, like drawing bar plot for each categorical feature, creating a table of summary statistics, plotting correlations, with a single command. While this speeds up the work significantly, it can be problematic for high-dimensional data and it does not take the advantage of AI tools for actual automatization. There is a lot of potential for intelligent data exploration (or model exploration) tools.