**The Problem**

There’s one thing about R that a lot of people have as their Top-of-Mind. That’s the black-and-white plot of iris dataset which is definitely a huge boring view of R. That’s boring because of aesthetics but also because it’s such a cliched example used over and over again. The other problem is finding the right set of dataset for the right set of problem you want to teach/learn/experiment. Let’s say you want to teach Time Series and that’s a case where your Spam / Ham Classification Dataset isn’t going to be of any use.

**Solution**

No more worries. That’s where fakir has arrived to help us. fakir is an R-package by [**Colin Fay**](https://www.programmingwithr.com/how-to-generate-meaningful-fake-data-for-learning-experimentation-and-teaching/%60fakir%60) (of Think-R) who’s been so good with his contributions to the R community.

**About fakir**

As in the documentation, The goal of fakir is to provide fake datasets that can be used to teach R.

**Installation and Loading**

fakir can be installed from Github (fakir isn’t available on CRAN yet)

# install.packages("devtools")

devtools::install\_github("ThinkR-open/fakir")

library(fakir)

**Use-case: Clickstream / Web Data**

Clickstream / Web Data is one thing a lot of organizations use in analytics these days but it’s hard to get your hand on some clickstream data since no company would prefer sharing theirs. There’s a sample Data on Google Analytics Test Account but that may not serve you any purpose in learning Data science in R or R’s ecosystem.

This is a typical case where fakir can help you

library(tidyverse)

fakir::fake\_visits() %>% head()

## # A tibble: 6 x 8

## timestamp year month day home about blog contact

##

## 1 2017-01-01 2017 1 1 352 176 521 NA

## 2 2017-01-02 2017 1 2 203 115 492 89

## 3 2017-01-03 2017 1 3 103 59 549 NA

## 4 2017-01-04 2017 1 4 484 113 633 331

## 5 2017-01-05 2017 1 5 438 138 423 227

## 6 2017-01-06 2017 1 6 NA 75 478 289

That’s how simple is to get a sample Clickstream (tidy) data with fakir. Another good thing to mention is, If you look at the fake\_visits() documentation, You’ll find it that there’s an argument that takes seed value which means, you are in control of randomizing the data and reproducing them.

fake\_visits(from = "2017-01-01", to = "2017-12-31", local = c("en\_US", "fr\_FR"),

seed = 2811) %>% head()

## # A tibble: 6 x 8

## timestamp year month day home about blog contact

##

## 1 2017-01-01 2017 1 1 352 176 521 NA

## 2 2017-01-02 2017 1 2 203 115 492 89

## 3 2017-01-03 2017 1 3 103 59 549 NA

## 4 2017-01-04 2017 1 4 484 113 633 331

## 5 2017-01-05 2017 1 5 438 138 423 227

## 6 2017-01-06 2017 1 6 NA 75 478 289

**Use-case: French Data**

Also, in the above usage of fake\_visits() function you might have noticed another attribute local which can help you select French data instead of English. In my personal opinion, This is crucial if you are on a mission of improving Data Literacy or Democratising Data Science.

fake\_ticket\_client(vol = 10, local = "fr\_FR") %>% head()

## # A tibble: 6 x 25

## ref num\_client prenom nom job age region id\_dpt departement

##

## 1 DOSS… 31 Const… Boul… 62 Pays … 44 Loire-Atla…

## 2 DOSS… 79 Martin Norm… Cons… 52 Alsace 67 Bas-Rhin

## 3 DOSS… 65 Phili… Géra… 28 Poito… 86 Vienne

## 4 DOSS… 77 Simon… Cour… Plom… 29 Île-d… 91

## 5 DOSS… 59 Rémy Dela… 18 Picar… 02 Aisne

## 6 DOSS… 141 Astrid Dumo… Ingé… 35 Nord-… 62 Pas-de-Cal…

## # … with 16 more variables: gestionnaire\_cb , nom\_complet ,

## # entry\_date , points\_fidelite , priorite\_encodee ,

## # priorite , timestamp , annee , mois , jour ,

## # pris\_en\_charge , pris\_en\_charge\_code , type ,

## # type\_encoded , etat , source\_appel

In the above example, We’ve used another function fake\_ticket\_client() of fakir that helps us in giving a typical ticket dataset (like the one you get from **ServiceNow** or **Zendesk**)

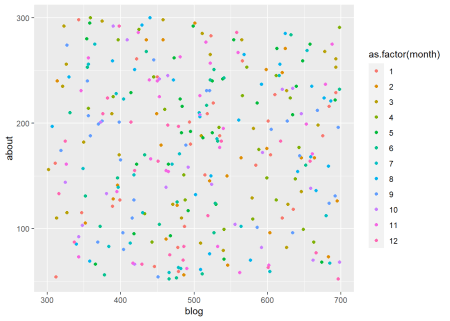
**Use-case: Scatter Plot**

So, the rant that I made at the start of this post about iris (Don’t mistake me: I’ve got huge respect for the scientists who created this dataset, it’s just that the wrong / over-usage of it which I don’t appreciate), Now we can overcome with fakir’s datasets.

fake\_visits() %>%

ggplot() + geom\_point(aes(blog,about, color = as.factor(month)))

## Warning: Removed 47 rows containing missing values (geom\_point).

  
(Perhaps, Not a good scatter plot to show Correlation but hey, you can teach scatter plot without plotting Petal Length and Sepal Length)

**Summary**

If you are in the business of teaching or likes experimenting and don’t want to use cliched datasets, fakir is a very nice package to get to know. As the author of fakir’s package mentions in the description, [charlatan](https://github.com/ropensci/charlatan) is another such R-package that helps in generating meaningful fake data.