Machine Learning for Factor Investing (Coqueret & Guida, 2021)

```
knitr::opts_chunk$set(message=FALSE, warning=FALSE)
```

Chapter 1

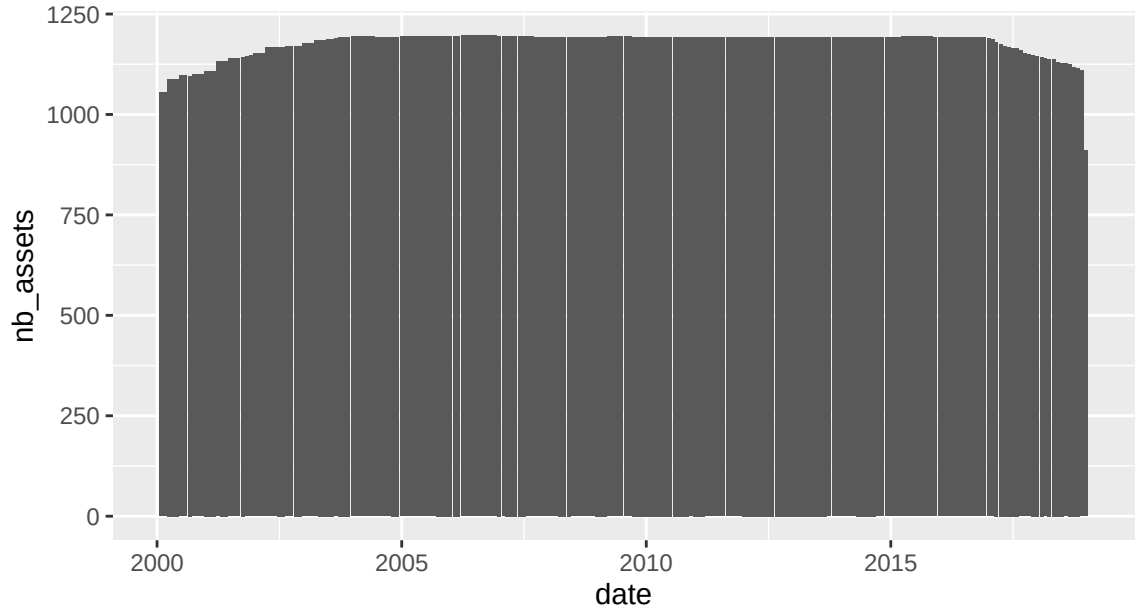
```
# load required libraries
library(tidyverse)
library(lubridate)
load('data_ml.RData')
```

```
data_ml <- data_ml %>%
    filter(date > "1999-12-31",
           date < "2019-01-01") %>%
    arrange(stock_id, date)
data_ml[1:6, 1:6]
```

```
## # A tibble: 6 x 6
##   stock_id date       Advt_12M_Usd Advt_3M_Usd Advt_6M_Usd Asset_Turnover
##      <int> <date>            <dbl>       <dbl>       <dbl>          <dbl>
## 1        1 2000-01-31         0.41        0.39        0.42           0.19
## 2        1 2000-02-29         0.41        0.39        0.4            0.19
## 3        1 2000-03-31         0.4         0.37        0.37           0.2
## 4        1 2000-04-30         0.39        0.36        0.37           0.2
## 5        1 2000-05-31         0.4         0.42        0.4            0.2
## 6        1 2000-06-30         0.41        0.47        0.42           0.21
```

Plot number of assets by date

```
data_ml %>%
    group_by(date) %>%
    summarize(nb_assets = stock_id %>% as.factor() %>% nlevels()) %>% # count number of assets
                  ggplot(aes(x = date, y = nb_assets)) + geom_col() +
                  coord_fixed(3)
```




```
features <- colnames(data_ml[3:95])
features_short <- c("Div_Yld", "Eps", "Mkt_Cap_12M_Usd", "Mom_11M_Usd", "Ocf", "Pb", "Vol1Y_Usd")
```

Create additional categorical labels.

```
data_ml <- data_ml %>%
            group_by(date) %>%
            mutate(R1M_Usd_C = R1M_Usd > median(R1M_Usd),
                   R12M_Usd_C = R1M_Usd > median(R12M_Usd)) %>%
            ungroup() %>%
            mutate_if(is.logical, as.factor)
```

Splitting train and test set.

```
separation_date <- as.Date("2014-01-15")
training <- filter(data_ml, date > separation_date)
testing_sample <- filter(data_ml, date <= separation_date)
```

Keep stocks with maximum number of points.

```
stock_ids <- levels(as.factor(data_ml$stock_id)) # a list of all stock_ids
stock_days <- data_ml %>% # compute the number of data points per stock
    group_by(stock_id) %>% summarize(nb = n())
stock_ids_short <- stock_ids[which(stock_days$nb == max(stock_days$nb))] # Stocks with full data
returns <- data_ml %>%
    filter(stock_id %in% stock_ids_short) %>% # 1. Filter the data
    dplyr::select(date, stock_id, R1M_Usd) %>% # 2. Keep returns along with dates and firms ID
    spread(key = stock_id, value = R1M_Usd) # 3. Put in matrix shape
```

Chapter 2