Datasets

```
load(file.path(data_dir, "data_ml.RData"))
```

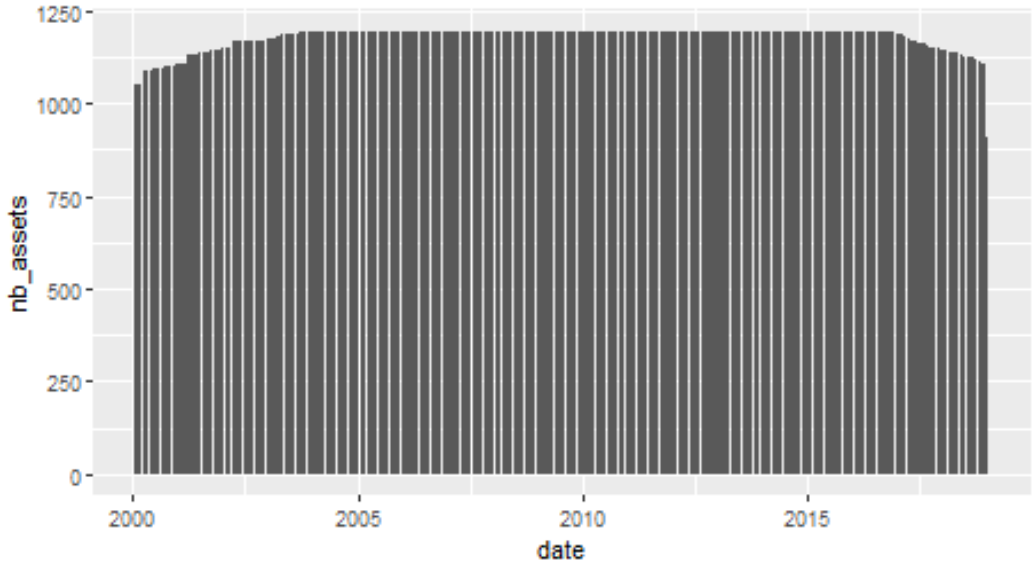
```
data_ml <- data_ml %>%
   filter(date > "1999-12-31",
          date < "2019-01-01") %>%
   arrange(stock_id, date)

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
```

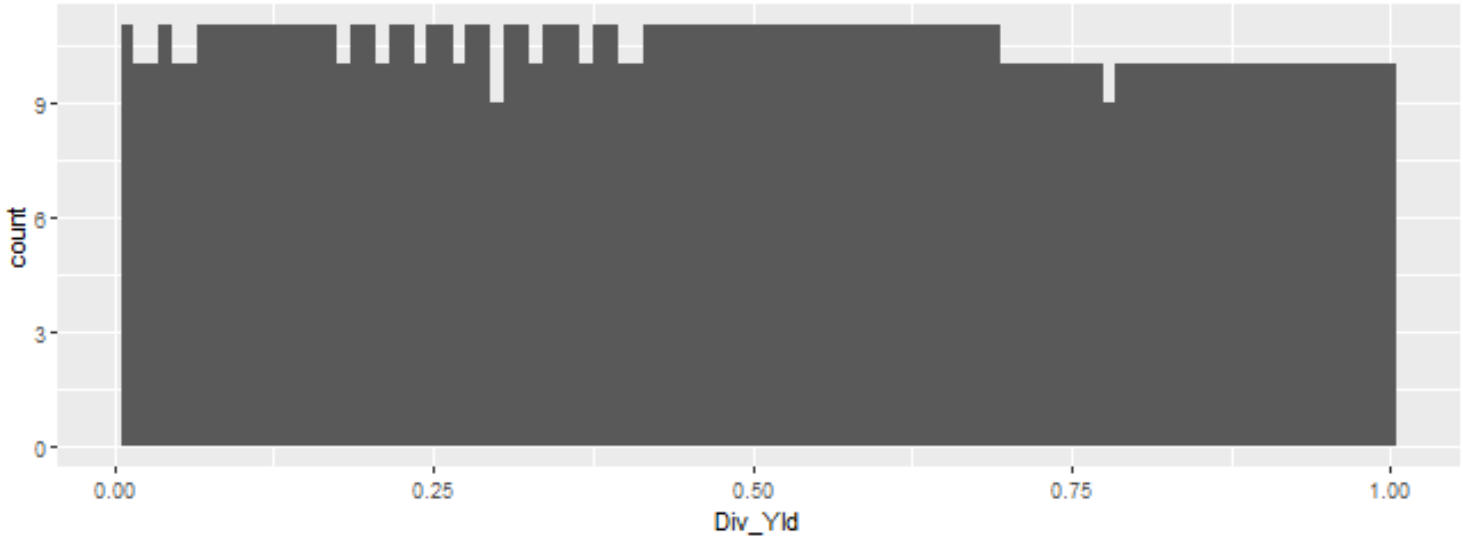
```
data_ml %>%
   group_by(date) %>%
   summarise(nb_assets = stock_id %>%
                as.factor() %>% nlevels()) %>%
   ggplot(aes(x = date, y = nb_assets)) +
   geom_col() +
   coord_fixed(3)
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```



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

```
data_ml %>%
   filter(date == "2000-02-29") %>%
   ggplot(aes(x = Div_Yld)) +
   geom_histogram(bins = 100) +
   coord_fixed(0.03)
```




```
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)
```

```
separation_date <- as_date("2014-01-15")

traning_sample <- filter(data_ml, date < separation_date)
testing_sample <- filter(data_ml, date > separation_date)

stock_ids <- levels(as.factor(data_ml$stock_id)) # list of all stock ids

stock_days <- data_ml %>%
   group_by(stock_id) %>%
   summarise(nb = n())
```

```
`summarise()` ungrouping output (override with `.groups` argument)
```

```
stock_ids_short <- stock_ids[which(stock_days$nb == max(stock_days$nb))] # keep only stocks wi

returns <- data_ml %>%
   filter(stock_id %in% stock_ids_short) %>%
   dplyr::select(date, stock_id, R1M_Usd) %>%
   spread(key = stock_id, value = R1M_Usd)
```