Challenge 03 – Solution

AUTHOR
Sebastian Sauer

PUBLISHED
November 22, 2025

1 Daten verstehen

1.1 Aufgaben

Datensatz pivotieren

Aufbauend auf dem Ergebnis der letzten Challenge:

1. *Pivotieren* Sie den Datensatz in das Langformat. Aber nehmen Sie ID-Variable `idvisit` vom Pivotieren aus; d.h. Sie nehmen alle Spalten bis auf `idvisit` in das Pivotieren auf. Es sollen also drei Spalten resultieren: `idvisit`, `name` und `value`.

2. Jetzt *pivotieren* Sie den Datensatz erneut in das Langformat. Aber dieses Mal verwenden Sie eine andere ID-Variablen, nämlich `fingerprint`. Ansonsten bleibt alles gleich. Es sollen also drei Spalten resultieren: `fingerprint`, `name` und `value`.

3. Jetzt *pivotieren* Sie den Datensatz erneut in das Langformat. Aber dieses Mal verwenden Sie zwei ID-Variaben: Die ID-Variable `idvisit` und auch `fingerprint`. Ansonsten bleibt alles gleich. Es sollen also *vier* Spalten resultieren: `idvisit`, `fingerprint`, `name` und `value`.

4. *Begrenzen* Sie beim Pivotieren die Spalten auf die Spaltentypen mit dem Namen vom Typ `subtitle`; d.h. Sie pivotieren nur diese genannten Spaltentypen. Es sollen also drei Spalten resultieren: `idvisit`, `name` und `value`.

5. Jetzt *begrenzen* Sie das Pivotieren wie in der vorherigen Aufgabe. Aber dieses Mal verwenden Sie die ID-Variable `fingerprint`. Es sollen also drei Spalten resultieren: `fingerprint`, `name` und `value`.

6. Jetzt begrenzen Sie das Pivotieren wie in der vorherigen Aufgabe. Aber dieses Mal verwenden Sie zwei ID-Variablen: `idvisit` und `fingerprint`. Es sollen also *vier* Spalten resultieren: `idvisit`, `fingerprint`, `name` und `value`.

7. Prüfen Sie, ob es stimmt, dass in der Spalte `name` die enthaltene Zahl die einzige Information ist. Anders gesagt: Außer der Zahl in den Werten `name` sind alle Teile der Werte konstant.

8. Die Spalten, die die Werte wie `actionDetails_0_subtitle` nennen Sie in `id` um. Aus den Werten (wie `actionDetails_0_subtitle`) extrahieren Sie die Zahl in der Mitte des Textes (den Rest des jeweiligen Spaltennamens löschen).

9. Erläutern Sie die Funktionen zum Lang-Pivotieren aus dem Tidyverse sowie aus einem anderen R-Paket, welches angibt, *schneller* zu sein als die Tidyverse-Funktion (also größere Datenmengen in der gleichen Zeit schafft). Geben Sie auch einen Faktor an, um diese Funktion angibt, schneller zu sein.

1.2 Setup

```
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.4     ✔ readr     2.1.6
✔ forcats   1.0.0     ✔ stringr   1.6.0
✔ ggplot2   4.0.1     ✔ tibble    3.3.0
✔ lubridate 1.9.4     ✔ tidyr     1.3.1
✔ purrr     1.2.0
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
#library(stringr)  # Strings verarbeiten
library(here)  # liest aktuelles Verzeichnis aus
```

```
here() starts at /home/sebastian-sauer/Dokumente/hans-hackathon2025
```

```
library(janitor)
```

```

Attache Paket: 'janitor'

Die folgenden Objekte sind maskiert von 'package:stats':

    chisq.test, fisher.test
```

```
#library(lubridate)  # Mit Zeitangaben arbeiten
library(tictoc)
library(data.table)
```

```

Attache Paket: 'data.table'

Das folgende Objekt ist maskiert 'package:tictoc':

    shift

Die folgenden Objekte sind maskiert von 'package:lubridate':

    hour, isoweek, mday, minute, month, quarter, second, wday, week,
    yday, year

Die folgenden Objekte sind maskiert von 'package:dplyr':

    between, first, last
```

```
Das folgende Objekt ist maskiert 'package:purrr':

    transpose
```

Wir laden den Datensatz, wie im letzten Schritt herausgegangen:

```
tic()
d_file_path <- "https://raw.githubusercontent.com/sebastiansauer/hans-hackathon2025/
d_input <- read_csv(d_file_path,
                    col_types = cols(.default = "c"))
```

```
New names:
• `` -> `...1`
```

```
toc()
```

```
0.498 sec elapsed
```

```
d_input_names_sanitized <-
  d_input |> clean_names()

names(d_input_names_sanitized) |>
  head(10)
```

```
 [1] "x1"                         "id_visit"
 [3] "fingerprint"                "action_details_0_subtitle"
 [5] "action_details_0_timestamp" "action_details_1_timestamp"
 [7] "action_details_1_subtitle"  "action_details_2_subtitle"
 [9] "action_details_2_timestamp" "action_details_3_timestamp"
```

```
d_input_names_sanitized |>
  select(1:20) |>
  glimpse()
```

```
Rows: 376
Columns: 20
$ x1                         <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9"…
$ id_visit                   <chr> "3294", "3292", "3293", "3291", "3290", "32…
$ fingerprint                <chr> "b60fd403ef2a6ad5", "f1f2268d3eb2725a", "95…
$ action_details_0_subtitle  <chr> "https://hswt.de&password=<ljse9wp0hps9y/lo…
$ action_details_0_timestamp <chr> "2025-07-07 23:16:05", "2025-07-07 22:07:02…
$ action_details_1_timestamp <chr> "2025-07-07 23:16:10", "2025-07-07 22:07:04…
$ action_details_1_subtitle  <chr> "Category: \"login', Action: \"success\"", …
$ action_details_2_subtitle  <chr> "https://hswt.de&password=<ljse9wp0hps9y/?e…
$ action_details_2_timestamp <chr> "2025-07-07 23:16:10", "2025-07-07 22:07:04…
$ action_details_3_timestamp <chr> "2025-07-07 23:16:25", "2025-07-07 22:07:12…
$ action_details_3_subtitle  <chr> "Category: \"click_videocard_search_semeste…
$ action_details_4_timestamp <chr> "2025-07-07 23:16:25", "2025-07-07 22:07:12…
$ action_details_4_subtitle  <chr> "Category: \"click_videocard_search_lecture…
$ action_details_5_subtitle  <chr> "Prof. Dr. Mircea Tric", "Category: \"stati…
$ action_details_5_timestamp <chr> "2025-07-07 23:16:35", "2025-07-07 22:07:14…
```

```
$ action_details_6_subtitle  <chr> "https://hswt.de&password=<ljse9wp0hps9y/se…
$ action_details_6_timestamp <chr> "2025-07-07 23:16:35", "2025-07-07 22:07:14…
$ action_details_7_timestamp <chr> "2025-07-07 23:16:57", "2025-07-07 22:07:14…
$ action_details_7_subtitle  <chr> "Category: \"click_videocard', Action: \"02…
$ action_details_8_subtitle  <chr> "https://hswt.de&password=<ljse9wp0hps9y/vi…
```

1.3 Lösungen

1.3.1 1 Pivotieren

Mit `id_visit` als ID:

```
tic()
d_long_idvisit <-
  d_input_names_sanitized |>
  select(-x1) |>
  pivot_longer(cols = -id_visit)
toc()
```

```
0.072 sec elapsed
```

```
glimpse(d_long_idvisit)
```

```
Rows: 377,504
Columns: 3
$ id_visit <chr> "3294", "3294", "3294", "3294", "3294", "3294", "3294", "3294…
$ name     <chr> "fingerprint", "action_details_0_subtitle", "action_details_0…
$ value    <chr> "b60fd403ef2a6ad5", "https://hswt.de&password=<ljse9wp0hps9y/…
```

1.3.2 Pivotieren mit `fingerprint` als ID

```
tic()
d_long_fingerprint <-
  d_input_names_sanitized |>
  select(-x1) |>
  pivot_longer(cols = -fingerprint)
toc()
```

```
0.047 sec elapsed
```

```
glimpse(d_long_fingerprint)
```

```
Rows: 377,504
Columns: 3
$ fingerprint <chr> "b60fd403ef2a6ad5", "b60fd403ef2a6ad5", "b60fd403ef2a6ad5"…
$ name        <chr> "id_visit", "action_details_0_subtitle", "action_details_0…
$ value       <chr> "3294", "https://hswt.de&password=<ljse9wp0hps9y/login?eva…
```

1.3.3 Pivotieren mit beiden ID-Variablen

```
tic()
d_long_idvisit_fingerprint <-
  d_input_names_sanitized |>
  select(-x1) |>
  pivot_longer(cols = -c(id_visit, fingerprint))
toc()
```

```
0.068 sec elapsed
```

```
glimpse(d_long_idvisit_fingerprint)
```

```
Rows: 377,128
Columns: 4
$ id_visit    <chr> "3294", "3294", "3294", "3294", "3294", "3294", "3294", "3…
$ fingerprint <chr> "b60fd403ef2a6ad5", "b60fd403ef2a6ad5", "b60fd403ef2a6ad5"…
$ name        <chr> "action_details_0_subtitle", "action_details_0_timestamp",…
$ value       <chr> "https://hswt.de&password=<ljse9wp0hps9y/login?evalId=none…
```

1.3.4 Pivotieren und Begrenzen Sie beim Pivotieren die Spalten auf die Spaltentypen - id_visit

```
d_input_names_sanitized_only_subtitle_cols <-
d_input_names_sanitized |>
  select(id_visit, contains("subtitle"))

tic()
d_long_idvisit_only_subtitle_cols_idvisit <-
  d_input_names_sanitized_only_subtitle_cols |>
  pivot_longer(cols = -id_visit)
toc()
```

```
0.023 sec elapsed
```

```
glimpse(d_long_idvisit_only_subtitle_cols_idvisit)
```

```
Rows: 188,000
Columns: 3
$ id_visit <chr> "3294", "3294", "3294", "3294", "3294", "3294", "3294", "3294…
$ name     <chr> "action_details_0_subtitle", "action_details_1_subtitle", "ac…
$ value    <chr> "https://hswt.de&password=<ljse9wp0hps9y/login?evalId=none&us…
```

BONUS:

Den Zeitverbrauch kann man sich mit `profvis` näher anschauen:

```
library(profvis)
profvis({
  d_long_idvisit_only_subtitle_cols_idvisit <-
    d_input_names_sanitized_only_subtitle_cols %>%
    pivot_longer(cols = -id_visit)
})
```

```
})
```

1.3.5 Pivotieren und Begrenzen Sie beim Pivotieren die Spalten auf die Spaltentypen - fingerprint

```
d_input_names_sanitized_only_subtitle_cols_fingerprint <-
d_input_names_sanitized |>
  select(fingerprint, contains("subtitle"))

tic()
d_long_idvisit_only_subtitle_cols_fingerprint <-
  d_input_names_sanitized_only_subtitle_cols_fingerprint |>
  pivot_longer(cols = -fingerprint)
toc()
```

```
0.03 sec elapsed
```

```
glimpse(d_long_idvisit_only_subtitle_cols_fingerprint)
```

```
Rows: 188,000
Columns: 3
$ fingerprint <chr> "b60fd403ef2a6ad5", "b60fd403ef2a6ad5", "b60fd403ef2a6ad5"…
$ name        <chr> "action_details_0_subtitle", "action_details_1_subtitle", …
$ value       <chr> "https://hswt.de&password=<ljse9wp0hps9y/login?evalId=none…
```

1.3.6 Pivotieren und Begrenzen Sie beim Pivotieren die Spalten auf die Spaltentypen - idvisit und fingerprint

```
d_input_names_sanitized_only_subtitle_cols_idvisit_fingerprint <-
d_input_names_sanitized |>
  select(id_visit, fingerprint, contains("subtitle"))

tic()
d_long_idvisit_only_subtitle_cols_idvisit_fingerprint <-
  d_input_names_sanitized_only_subtitle_cols_idvisit_fingerprint |>
  pivot_longer(cols = -c(id_visit, fingerprint))
toc()
```

```
0.029 sec elapsed
```

```
glimpse(d_long_idvisit_only_subtitle_cols_idvisit_fingerprint)
```

```
Rows: 188,000
Columns: 4
$ id_visit    <chr> "3294", "3294", "3294", "3294", "3294", "3294", "3294", "3…
$ fingerprint <chr> "b60fd403ef2a6ad5", "b60fd403ef2a6ad5", "b60fd403ef2a6ad5"…
$ name        <chr> "action_details_0_subtitle", "action_details_1_subtitle", …
$ value       <chr> "https://hswt.de&password=<ljse9wp0hps9y/login?evalId=none…
```

1.3.7 rüfen Sie, ob es stimmt, dass in der Spalte …

```
d_long_idvisit_only_subtitle_cols_idvisit_fingerprint %>%
  pull(name) %>%
  str_remove("\\d+") %>%        # remove digits from each string
  unique() %>%                  # keep only unique values
  length()                      # count them
```

```
[1] 1
```

Ja, ist konstant ohne die Zahlen in der Mitte

1.3.8 Zahl extrahieren aus `name`

```
d_long_idvisit_only_subtitle_cols_idvisit_fingerprint |>
  mutate(id = str_extract(name, "\\d+")) |>
  select(-name)
```

```
# A tibble: 188,000 × 4
   id_visit fingerprint      value                                          id
   <chr>    <chr>            <chr>                                          <chr>
 1 3294     b60fd403ef2a6ad5 "https://hswt.de&password=<ljse9wp0hps9y/log… 0
 2 3294     b60fd403ef2a6ad5 "Category: \"login', Action: \"success\""     1
 3 3294     b60fd403ef2a6ad5 "https://hswt.de&password=<ljse9wp0hps9y/?ev… 2
 4 3294     b60fd403ef2a6ad5 "Category: \"click_videocard_search_semester… 3
 5 3294     b60fd403ef2a6ad5 "Category: \"click_videocard_search_lecturer… 4
 6 3294     b60fd403ef2a6ad5 "Prof. Dr. Mircea Tric"                        5
 7 3294     b60fd403ef2a6ad5 "https://hswt.de&password=<ljse9wp0hps9y/sea… 6
 8 3294     b60fd403ef2a6ad5 "Category: \"click_videocard', Action: \"02_… 7
 9 3294     b60fd403ef2a6ad5 "https://hswt.de&password=<ljse9wp0hps9y/vid… 8
10 3294     b60fd403ef2a6ad5 "Category: \"videoplayer_click', Action: \"p… 9
# ℹ 187,990 more rows
```

1.3.9 Schneller als `pivot_longer`

`data.table` ist eines der bekanntesten R-Pakete. Es glänzt durch Geschwindigkeit.

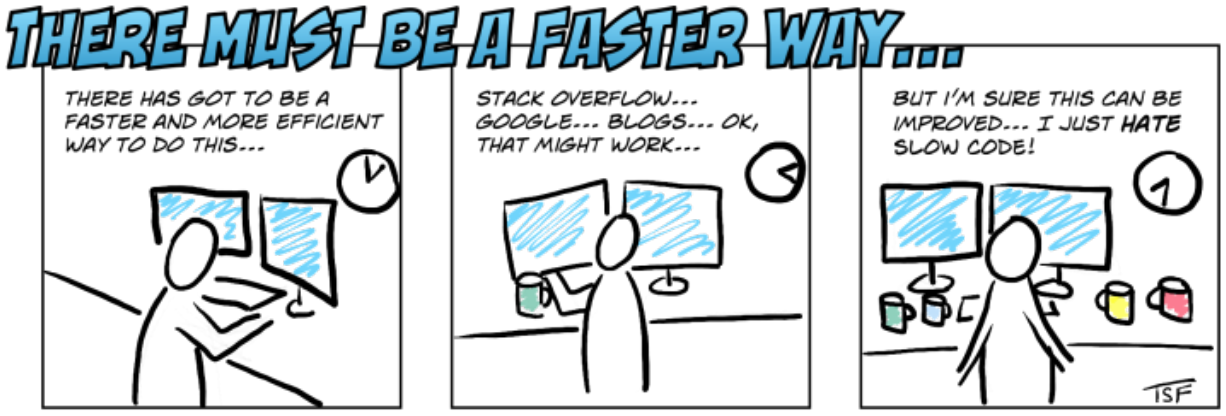
```
vars_to_pivot <-
  d_input_names_sanitized %>%
  select(contains("action_details_")) %>%
  names()

DT <- as.data.table(d_input_names_sanitized)

tic()
out <- melt(DT,
    id.vars = c("id_visit", "fingerprint"),
    measure.vars = vars_to_pivot,
    variable.name = "name",
    value.name = "value"
  )
toc()
```

```
0.032 sec elapsed
```

1.3.10 Bonus: Wer ist schneller?



`data.table` ist deutlich schneller als `tidyverse`:

- https://duckdblabs.github.io/db-benchmark/
- https://towardsdatascience.com/data-table-speed-with-dplyr-syntax-yes-we-can-51ef9aaed585/
- https://timfarewell.co.uk/is-data-table-or-dplyr-faster-at-summarising-data/
- https://codepointtech.com/scale-tidyverse-to-big-data-master-dtplyr-for-dplyr-data-table/

1.4 Outro

Als RDS-Datei:

```
tic()
write_rds(d_long_idvisit_only_subtitle_cols_idvisit,
          paste0(here(),"/objects/d_long_idvisit_only_subtitle_cols_idvisit.rds"))
toc()
```

```
0.193 sec elapsed
```

```
write_rds(d_long_idvisit_only_subtitle_cols_idvisit_fingerprint,
          paste0(here(),"/objects/d_long_idvisit_only_subtitle_cols_idvisit_fingerpr
```

Als Text-Datei:

```
tic()
write_csv(d_long_idvisit_only_subtitle_cols_idvisit,
          paste0(here(),"/objects/d_long_idvisit_only_subtitle_cols_idvisit.csv"))
toc()
```

```
0.155 sec elapsed
```

```
write_csv(d_long_idvisit_only_subtitle_cols_idvisit_fingerprint,
          paste0(here(),"/objects/d_long_idvisit_only_subtitle_cols_idvisit_fingerpr
```

2 SessionInfo

```
sessionInfo()
```

```
R version 4.5.1 (2025-06-13)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 25.10

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.12.1
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.1;  LAPACK version 3.12.0

locale:
 [1] LC_CTYPE=de_DE.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=de_DE.UTF-8        LC_COLLATE=de_DE.UTF-8
 [5] LC_MONETARY=de_DE.UTF-8    LC_MESSAGES=de_DE.UTF-8
 [7] LC_PAPER=de_DE.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=de_DE.UTF-8 LC_IDENTIFICATION=C

time zone: Europe/Berlin
tzcode source: system (glibc)

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] data.table_1.17.8 tictoc_1.2.1      janitor_2.2.1     here_1.0.1
 [5] lubridate_1.9.4   forcats_1.0.0     stringr_1.6.0     dplyr_1.1.4
 [9] purrr_1.2.0       readr_2.1.6       tidyr_1.3.1       tibble_3.3.0
[13] ggplot2_4.0.1     tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] utf8_1.2.6          generics_0.1.4     stringi_1.8.7      hms_1.1.3
 [5] digest_0.6.39       magrittr_2.0.4     evaluate_1.0.5     grid_4.5.1
 [9] timechange_0.3.0    RColorBrewer_1.1-3 fastmap_1.2.0      rprojroot_2.1.1
[13] jsonlite_2.0.0      scales_1.4.0       cli_3.6.5          rlang_1.1.6
[17] crayon_1.5.3        bit64_4.6.0-1      withr_3.0.2        yaml_2.3.10
[21] parallel_4.5.1      tools_4.5.1        tzdb_0.5.0         curl_7.0.0
[25] vctrs_0.6.5         R6_2.6.1           lifecycle_1.0.4    snakecase_0.11.1
[29] htmlwidgets_1.6.4   bit_4.6.0          vroom_1.6.5        pkgconfig_2.0.3
[33] pillar_1.11.1       gtable_0.3.6       glue_1.8.0         xfun_0.54
[37] tidyselect_1.2.1    rstudioapi_0.17.1  knitr_1.50         dichromat_2.0-0.1
[41] farver_2.1.2        htmltools_0.5.8.1  rmarkdown_2.30     compiler_4.5.1
[45] S7_0.2.1
```