Load and clean Excel files using tidyxl and unpivotr part 1

Gladys Wojciechowska

22 June 2021

```
# * Load libraries -----
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.1     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(unpivotr)
```

```
##
## Attaching package: 'unpivotr'
```

```
## The following objects are masked from 'package:tidyr':
##
##     pack, unpack
```

```
library(tidyxl)

# * Load data using tidyxl::xlsx_cells-----

test <- xlsx_cells("sample_data.xlsx")

head(test)
```

```
## # A tibble: 6 x 21
##   sheet    address   row   col is_blank data_type error logical numeric
##   <chr>    <chr>   <int> <int> <lgl>    <chr>     <chr> <lgl>     <dbl>
## 1 Sample 1 A1          1     1 FALSE    character <NA>  NA           NA
## 2 Sample 1 B1          1     2 FALSE    character <NA>  NA           NA
## 3 Sample 1 C1          1     3 FALSE    character <NA>  NA           NA
## 4 Sample 1 D1          1     4 TRUE     blank     <NA>  NA           NA
## 5 Sample 1 E1          1     5 TRUE     blank     <NA>  NA           NA
## 6 Sample 1 F1          1     6 TRUE     blank     <NA>  NA           NA
## # ... with 12 more variables: date <dttm>, character <chr>,
## #   character_formatted <list>, formula <chr>, is_array <lgl>,
## #   formula_ref <chr>, formula_group <int>, comment <chr>, height <dbl>,
## #   width <dbl>, style_format <chr>, local_format_id <int>
```

```
tail(test)
```

```
## # A tibble: 6 x 21
##   sheet    address   row   col is_blank data_type error logical numeric
##   <chr>    <chr>   <int> <int> <lgl>    <chr>     <chr> <lgl>     <dbl>
## 1 Sample 2 F13        13     6 FALSE    numeric   <NA>  NA         11.4
## 2 Sample 2 G13        13     7 FALSE    numeric   <NA>  NA         95
## 3 Sample 2 H13        13     8 FALSE    character <NA>  NA         NA
## 4 Sample 2 H14        14     8 TRUE     blank     <NA>  NA         NA
## 5 Sample 2 H15        15     8 TRUE     blank     <NA>  NA         NA
## 6 Sample 2 H16        16     8 TRUE     blank     <NA>  NA         NA
## # ... with 12 more variables: date <dttm>, character <chr>,
## #   character_formatted <list>, formula <chr>, is_array <lgl>,
## #   formula_ref <chr>, formula_group <int>, comment <chr>, height <dbl>,
## #   width <dbl>, style_format <chr>, local_format_id <int>
```

```
# How many Excel sheets do we have?

xlsx_sheet_names("sample_data.xlsx")
```

```
## [1] "Sample 1" "Sample 2"
```

```
# Load the first sheet using two options

test_1a <- xlsx_cells("sample_data.xlsx", sheets = 1)
test_1b <- xlsx_cells("sample_data.xlsx", sheets = "Sample 1")

identical(test_1a, test_1b)
```

```
## [1] TRUE
```

```
# * Explore the data (First sheet) ----
data_1 <- xlsx_cells("sample_data.xlsx", sheets = 1)
print(data_1 %>% filter(row == 9), width = Inf)
```

```
## # A tibble: 8 x 21
##   sheet    address   row   col is_blank data_type error logical numeric
##   <chr>    <chr>   <int> <int> <lgl>    <chr>     <chr> <lgl>     <dbl>
## 1 Sample 1 A9          9     1 FALSE    numeric   <NA>  NA          6
## 2 Sample 1 B9          9     2 FALSE    character <NA>  NA         NA
## 3 Sample 1 C9          9     3 FALSE    character <NA>  NA         NA
## 4 Sample 1 D9          9     4 FALSE    numeric   <NA>  NA         97
## 5 Sample 1 E9          9     5 TRUE     blank     <NA>  NA         NA
## 6 Sample 1 F9          9     6 FALSE    character <NA>  NA         NA
## 7 Sample 1 G9          9     7 FALSE    numeric   <NA>  NA         10.5
## 8 Sample 1 H9          9     8 FALSE    numeric   <NA>  NA         97
##   date                character character_formatted    formula is_array
##   <dttm>              <chr>     <list>                 <chr>   <lgl>
## 1 NA                  <NA>      <NULL>                 <NA>    FALSE
## 2 NA                  Zacma     <tibble[,14] [1 x 14]> <NA>    FALSE
## 3 NA                  N/A       <tibble[,14] [1 x 14]> <NA>    FALSE
## 4 NA                  <NA>      <NULL>                 <NA>    FALSE
## 5 NA                  <NA>      <NULL>                 <NA>    FALSE
## 6 NA                  n/a       <tibble[,14] [1 x 14]> <NA>    FALSE
## 7 NA                  <NA>      <NULL>                 <NA>    FALSE
## 8 NA                  <NA>      <NULL>                 <NA>    FALSE
##   formula_ref formula_group comment height width style_format local_format_id
##   <chr>               <int> <chr>    <dbl> <dbl> <chr>                  <int>
## 1 <NA>                   NA <NA>        13  8.73 Normal                    10
## 2 <NA>                   NA <NA>        13 14.5  Normal                     3
## 3 <NA>                   NA <NA>        13  8.73 Normal                    13
## 4 <NA>                   NA <NA>        13  8.73 Normal                     9
## 5 <NA>                   NA <NA>        13  8.73 Normal                     9
## 6 <NA>                   NA <NA>        13  8.73 Normal                    15
## 7 <NA>                   NA <NA>        13  8.73 Normal                     9
## 8 <NA>                   NA <NA>        13  8.73 Normal                    11
```

```
names(data_1)
```

```
##  [1] "sheet"               "address"             "row"
##  [4] "col"                 "is_blank"            "data_type"
##  [7] "error"               "logical"             "numeric"
## [10] "date"                "character"           "character_formatted"
## [13] "formula"             "is_array"            "formula_ref"
## [16] "formula_group"       "comment"             "height"
## [19] "width"               "style_format"        "local_format_id"
```

```
# what kind of data types do we have in this sheet?

table(data_1$data_type)
```

```
##
##     blank character   numeric
##        22        25        63
```

```
# The selected variables from this sheet
data_1 %>%
  select(row, col, data_type, numeric, character, local_format_id)
```

```
## # A tibble: 110 x 6
##      row   col data_type numeric character     local_format_id
##    <int> <int> <chr>       <dbl> <chr>                   <int>
##  1     1     1 character      NA ID                          4
##  2     1     2 character      NA History                    17
##  3     1     3 character      NA Lab test                   16
##  4     1     4 blank          NA <NA>                       16
##  5     1     5 blank          NA <NA>                       16
##  6     1     6 blank          NA <NA>                       16
##  7     1     7 blank          NA <NA>                       16
##  8     1     8 blank          NA <NA>                       16
##  9     2     1 blank          NA <NA>                        4
## 10     2     2 character      NA Comorbidities               5
## # ... with 100 more rows
```

```
# Move header names to a dedicated column using unpivotr::behead -----

# First beheading
data_1 %>%
  select(row, col, data_type, numeric, character, local_format_id) %>%
  behead("up", header_1)
```

```
## # A tibble: 102 x 7
##      row   col data_type numeric character           local_format_id header_1
##    <int> <int> <chr>       <dbl> <chr>                         <int> <chr>
##  1     2     1 blank          NA <NA>                              4 ID
##  2     2     2 character      NA Comorbidities                     5 History
##  3     2     3 character      NA Biochemistry Time 1               6 Lab test
##  4     2     4 blank          NA <NA>                              6 <NA>
##  5     2     5 blank          NA <NA>                              6 <NA>
##  6     2     6 blank          NA <NA>                              6 <NA>
##  7     2     7 blank          NA <NA>                              6 <NA>
##  8     2     8 blank          NA <NA>                              6 <NA>
##  9     3     1 blank          NA <NA>                              4 ID
## 10     3     2 blank          NA <NA>                              5 History
## # ... with 92 more rows
```

```
# Second beheading

data_1 %>%
  select(row, col, data_type, numeric, character, local_format_id) %>%
  behead("up", header_1) %>%
  behead("up", header_2) %>%
  print(width = Inf)
```

```
## # A tibble: 94 x 8
##      row   col data_type numeric character   local_format_id header_1
##    <int> <int> <chr>       <dbl> <chr>                 <int> <chr>
##  1     3     1 blank          NA <NA>                      4 ID
##  2     3     2 blank          NA <NA>                      5 History
##  3     3     3 character      NA Test 1                    7 Lab test
##  4     3     4 character      NA Test 2                    7 <NA>
##  5     3     5 character      NA Test 3                    7 <NA>
##  6     3     6 character      NA Test 4                    7 <NA>
##  7     3     7 character      NA Test 5                    7 <NA>
##  8     3     8 character      NA Test 6                    7 <NA>
##  9     4     1 numeric         1 <NA>                      8 ID
## 10     4     2 character      NA Rak zoladka               3 History
##    header_2
##    <chr>
##  1 <NA>
##  2 Comorbidities
##  3 Biochemistry Time 1
##  4 <NA>
##  5 <NA>
##  6 <NA>
##  7 <NA>
##  8 <NA>
##  9 <NA>
## 10 Comorbidities
## # ... with 84 more rows
```

```
# Last beheading

data_1 %>%
  select(row, col, data_type, numeric, character, local_format_id) %>%
  behead("up", header_1) %>%
  behead("up", header_2) %>%
  behead("up", header_3) %>%
  print(width = Inf)
```

```
## # A tibble: 86 x 9
##      row   col data_type numeric character   local_format_id header_1
##    <int> <int> <chr>       <dbl> <chr>                 <int> <chr>
##  1     4     1 numeric       1   <NA>                      8 ID
##  2     4     2 character    NA   Rak zoladka               3 History
##  3     4     3 numeric      11.0 <NA>                      9 Lab test
##  4     4     4 numeric      85   <NA>                      9 <NA>
##  5     4     5 numeric      12   <NA>                      9 <NA>
##  6     4     6 numeric     111   <NA>                     18 <NA>
##  7     4     7 numeric      10.0 <NA>                      9 <NA>
##  8     4     8 numeric      85   <NA>                     11 <NA>
##  9     5     1 numeric       2   <NA>                      8 ID
## 10     5     2 numeric       1   <NA>                      3 History
##    header_2            header_3
##    <chr>               <chr>
##  1 <NA>                <NA>
##  2 Comorbidities       <NA>
##  3 Biochemistry Time 1 Test 1
##  4 <NA>                Test 2
##  5 <NA>                Test 3
##  6 <NA>                Test 4
##  7 <NA>                Test 5
##  8 <NA>                Test 6
##  9 <NA>                <NA>
## 10 Comorbidities       <NA>
## # ... with 76 more rows
```

```
# Create a header column with the proper header names, then spatter

data_1 %>%
  select(row, col, data_type, numeric, character, local_format_id) %>%
  behead("up", header_1) %>%
  behead("up", header_2) %>%
  behead("up", header_3) %>%
  mutate(header = case_when(header_1 == "ID" ~ "id",
                            header_1 == "History" ~ "history",
                            header_3 == "Test 1" ~ "biochem_1",
                            header_3 == "Test 2" ~ "biochem_2",
                            header_3 == "Test 3" ~ "biochem_3",
                            header_3 == "Test 4" ~ "biochem_4",
                            header_3 == "Test 5" ~ "biochem_5",
                            header_3 == "Test 6" ~ "biochem_6")) %>%
  print(width = Inf)
```

```
## # A tibble: 86 x 10
##      row   col data_type numeric character   local_format_id header_1
##    <int> <int> <chr>       <dbl> <chr>                 <int> <chr>
##  1     4     1 numeric       1   <NA>                      8 ID
##  2     4     2 character    NA   Rak zoladka               3 History
##  3     4     3 numeric      11.0 <NA>                      9 Lab test
##  4     4     4 numeric      85   <NA>                      9 <NA>
##  5     4     5 numeric      12   <NA>                      9 <NA>
##  6     4     6 numeric     111   <NA>                     18 <NA>
##  7     4     7 numeric      10.0 <NA>                      9 <NA>
##  8     4     8 numeric      85   <NA>                     11 <NA>
##  9     5     1 numeric       2   <NA>                      8 ID
## 10     5     2 numeric       1   <NA>                      3 History
##    header_2            header_3 header
##    <chr>               <chr>    <chr>
##  1 <NA>                <NA>     id
##  2 Comorbidities       <NA>     history
##  3 Biochemistry Time 1 Test 1   biochem_1
##  4 <NA>                Test 2   biochem_2
##  5 <NA>                Test 3   biochem_3
##  6 <NA>                Test 4   biochem_4
##  7 <NA>                Test 5   biochem_5
##  8 <NA>                Test 6   biochem_6
##  9 <NA>                <NA>     id
## 10 Comorbidities       <NA>     history
## # ... with 76 more rows
```

```
data_1 <- data_1 %>%
  select(row, col, data_type, numeric, character, local_format_id) %>%
  behead("up", header_1) %>%
  behead("up", header_2) %>%
  behead("up", header_3) %>%
  mutate(header = case_when(header_1 == "ID" ~ "id",
                            header_1 == "History" ~ "history",
                            header_3 == "Test 1" ~ "biochem_1",
                            header_3 == "Test 2" ~ "biochem_2",
                            header_3 == "Test 3" ~ "biochem_3",
                            header_3 == "Test 4" ~ "biochem_4",
                            header_3 == "Test 5" ~ "biochem_5",
                            header_3 == "Test 6" ~ "biochem_6")) %>%
  select(row, data_type, numeric, character, header) %>%
  spatter(header) %>%
  select(row, id, history, everything())

# The clean data frame! Save as csv.

print(data_1, width = Inf)
```

```
## # A tibble: 13 x 9
##      row    id history                 biochem_1        biochem_2 biochem_3
##    <int> <dbl> <chr>                   <chr>            <chr>         <dbl>
##   1     4     1 Rak zoladka             11.0322569924589 85               12
##   2     5     2 1                       10.4969141076758 179              10
##   3     6     3 Rak pluc                10.0514039930496 brak             28
##   4     7     4 0                       10.9305472190151 107              13
##   5     8     5 Rak pecherza moczowego N/A              174              21
##   6     9     6 Zacma                   N/A              97               NA
##   7    10     7 2                       10.7651254424628 172              NA
##   8    11     8 Cukrzyca                10.0250142655581 157              25
##   9    12     9 0                       10.0354453288257 brak             17
## 10    13    10 1                       10.0275722001274 brak             NA
## 11    14    NA <NA>                    <NA>             <NA>             NA
## 12    15    NA <NA>                    <NA>             <NA>             NA
## 13    16    NA <NA>                    <NA>             <NA>             NA
##    biochem_4 biochem_5        biochem_6
##    <chr>     <chr>                <dbl>
##   1 111       10.0046566810737        85
##   2 140       10.0668101039554       179
##   3 154       NA                     119
##   4 103       10.1719369238991       107
##   5 23        10.0715875417757       174
##   6 n/a       10.5153564940351        97
##   7 75        10.5812797830786       172
##   8 103       NA                     157
##   9 179       10.1464211583871        85
## 10 n/a       11.2248382695051       104
## 11 <NA>      <NA>                     NA
## 12 <NA>      <NA>                     NA
## 13 <NA>      <NA>                     NA
```

```
write_csv(data_1, "data_1_part1.csv")
```