Initial data cleanup (WIP)

Virginia Scarlett

2025-05-23

Overview

This script reads in the raw survey data from Qualtrics and splits the data into multiple files. It assumes the data were exported from Qualtrics using 'More Options' > 'Split multi-value fields into columns'.

It also assumes the data file is called 'raw_survey_data.tsv'.

I'm storing the path to my data folder in .Renviron like so: DATA_PATH = "/Path/to/data/folder" Because the data contain personally identifiable information, I'm grabbing them from elsewhere in my filesystem, outside of this project directory.

Import packages and utilities

```
project_root <- here::here() # requires that you be in project directory if using VS Code (i
print(paste0("Project root: ",project_root))
```

```
[1] "Project root: /Users/virginiascarlett/ospo-survey-analysis"
```

```
suppressMessages(source(file.path(project_root, "scripts/packages.R")))
suppressMessages(source(file.path(project_root, "scripts/utils.R")))
```

Define functions specific to this script

```
write_subset_of_data <- function(df, filen) {
  write.table(df,
    file.path(Sys.getenv("DATA_PATH"), filen),
    quote = FALSE,
    row.names = FALSE,
    sep = "\t"
  )
}
```

Load data

Open the raw survey data from Qualtrics. N.B. Qualtrics exports in UTF-16.

```
data <- load_qualtrics_data("raw_survey_data.tsv", fileEncoding = "utf-16")
```

Inspect data

As mentioned above in Overview, we used Qualtrics' 'Split multi-value fields into columns' function when we exported the data from Qualtrics, which causes every response field to be its own column. So if a multiple-choice question has 4 possible responses, it will have 4 columns in the dataframe.
Here are the dimensions of the data:

```
dim(data)
```

```
[1] 334 121
```

Here are the columns, as a vector:

```
names(data)
```

```
  [1] "StartDate"                 "EndDate"
  [3] "Status"                    "Progress"
  [5] "Duration (in seconds)"     "Finished"
  [7] "RecordedDate"              "ResponseId"
  [9] "DistributionChannel"       "UserLanguage"
 [11] "Q_RecaptchaScore"          "consent_form_2"
 [13] "campus"                    "importance_opensrc_1"
 [15] "importance_opensrc_2"      "importance_opensrc_3"
```

```
 [17] "importance_opensrc_4"        "importance_opensrc_5"
 [19] "contributor_status_1"        "contributor_status_2"
 [21] "contributor_role_1"          "contributor_role_6"
 [23] "contributor_role_7"          "contributor_role_12"
 [25] "contributor_role_13"         "contributor_role_14"
 [27] "contributor_role_8"          "contributor_role_15"
 [29] "contributor_role_10"         "contributor_role_11"
 [31] "contributor_role_11_TEXT"    "project_size_1"
 [33] "project_size_2"              "project_size_3"
 [35] "motivations_1"               "motivations_4"
 [37] "motivations_6"               "motivations_7"
 [39] "motivations_8"               "motivations_9"
 [41] "motivations_10"              "motivations_11"
 [43] "motivations_11_TEXT"         "project_types_7"
 [45] "project_types_1"             "project_types_4"
 [47] "project_types_5"             "project_types_6"
 [49] "project_types_3"             "project_types_2"
 [51] "project_types_2_TEXT"        "hosting_services_1"
 [53] "hosting_services_4"          "hosting_services_17"
 [55] "hosting_services_12"         "hosting_services_13"
 [57] "hosting_services_6"          "hosting_services_5"
 [59] "hosting_services_7"          "hosting_services_8"
 [61] "hosting_services_15"         "hosting_services_14"
 [63] "hosting_services_9"          "hosting_services_19"
 [65] "hosting_services_16"         "hosting_services_11"
 [67] "hosting_services_18"         "hosting_services_20"
 [69] "hosting_services_10"         "hosting_services_10_TEXT"
 [71] "challenges_1"                "challenges_2"
 [73] "challenges_3"                "challenges_4"
 [75] "challenges_5"                "challenges_6"
 [77] "challenges_7"                "challenges_8"
 [79] "challenges_9"                "challenges_10"
 [81] "challenges_11"               "challenges_12"
 [83] "challenges_13"               "challenges_14"
 [85] "solution_offerings_1"        "solution_offerings_2"
 [87] "solution_offerings_3"        "solution_offerings_4"
 [89] "solution_offerings_5"        "solution_offerings_6"
 [91] "solution_offerings_7"        "solution_offerings_8"
 [93] "solution_offerings_9"        "solution_offerings_10"
 [95] "solution_offerings_11"       "solution_offerings_12"
 [97] "favorite_solution"           "final_thoughts"
 [99] "usernames"                   "orb_followup_yes_1"
[101] "orb_followup_email"          "future_contributors_4"
```

```
[103] "future_contributors_8"       "future_contributors_1"
[105] "future_contributors_24"      "future_contributors_7"
[107] "future_contributors_23"      "future_contributors_11"
[109] "future_contributors_10"      "future_contributors_9"
[111] "future_contributors_22"      "future_contributors_12"
[113] "future_contributors_12_TEXT" "job_category"
[115] "field_of_study"              "subfield"
[117] "staff_categories"            "staff_categories_13_TEXT"
[119] "stay_in_touch_boxes_1"       "stay_in_touch_boxes_2"
[121] "stay_in_touch_email"
```

The column names are based on the question names in Qualtrics (which I chose). I don't know why the numbers at the end are kind of arbitrary for certain questions. I'm pretty sure Qualtrics automatically added that "_TEXT" suffix to the free-response question columns.