Data cleanup, part 1

In this script, I take the raw survey data downloaded from Qualtrics, and break it up into smaller tables that will be easier to work with. This script writes three data files. One of them, all_quant.tsv, will need further cleanup (see data_cleanup_part2.qmd).

This script assumes that when the raw data were downloaded from Qualtrics, tsv format was selected, as was 'More Options' > 'Split multi-value fields into columns'.

Make sure your data path is set in ~/.Renviron like so:
`DATA_PATH = "/Path/to/data/folder"`
If you don't want to do edit your global .Renviron file, you can edit the paths in `scripts/utils.R`.

Input:
raw_survey_data.tsv (downloaded from Qualtrics)

Output:
pii.tsv (contains emails, GitHub usernames, and contact preferences)
qual_responses.tsv (contains responses to free text boxes)
all_quant.tsv (contains responses to matrix and multiple-choice Qs)

Load packages

```
project_root <- here::here() # requires that you be somewhere in the
# project directory (not above it)
suppressMessages(source(file.path(project_root, "scripts/packages.R")))
# functions and objects used across scripts
suppressMessages(source(file.path(project_root, "scripts/utils.R")))
```

Functions

See `scripts/utils.R` for the `write_df_to_file` function.

Load raw data

```
# N.B. Qualtrics exports in UTF-16
data <- load_qualtrics_data("raw_survey_data.tsv", fileEncoding = "utf-16")
```

Drop unnecessary rows and columns

Drop rows where the “Finished” column is not “True”. This excludes unfinished survey responses and has the added benefit of removing those first two junk rows that Qualtrics generated and that we don’t care about.

```
data <- data %>% filter(Finished == "True")
```

Qualtrics also adds a bunch of junk columns at the beginning that we don’t care about, e.g. StartDate, EndDate, Duration. Drop these.

```
data <- data %>% select(consent_form_2:last_col())
# ^Not sure why qualtrics names this column "consent_form_2"
# instead of just "consent_form" but whatever.
```

Qualtrics also arranges columns in a sort of arbitrary order; let’s reorder them. Note, this command sorts alphabetically so questions are no longer in survey order.

```
data <- data %>% select(mixedsort(names(.)))
```

Write personally identifiable information (PII) to a file

```
pii_cols <- c(
  "usernames",
  "orb_followup_yes_1",
  "orb_followup_email",
  "stay_in_touch_boxes_1",
  "stay_in_touch_boxes_2",
  "stay_in_touch_email"
)

pii <- data %>% select(all_of(pii_cols))

write_df_to_file(pii, "pii.tsv")

data <- data %>% select(-all_of(pii_cols))
```

Write qualitative responses to a file

Note that I am curating the data a bit here. Some people hit 'return' in their text responses, which wreaked havoc on my parsing. So I am replacing tabs and newlines with a space.

PARTICIPANTS WHO DIVULGED PERSONAL INFO THAT MUST BE MANUALLY CLEANED UP AFTER RUNNING THIS SCRIPT: 75 114

```
qual_cols <- c(
  "final_thoughts",
  "subfield"
)

qual <- data %>% select(ends_with("_TEXT"), all_of(qual_cols))

qual_clean <- qual %>%
  mutate(across(everything(),
                ~ str_replace_all(as.character(.x), "[\\t\\n]", " ")))

write_df_to_file(qual_clean, "qual_responses.tsv")

data <- data %>% select(-ends_with("_TEXT"), -all_of(qual_cols))
```

Save quantitative data

```
write_df_to_file(data, "all_quant.tsv")
```