Module 1: Anatomy of `merged.csv`

April 23, 2021

Contents

The PNS documentation (e.g., see https://github.com/jamieyap/PNS) contains information on study design, how the data collected differed from the intended design, and how such deviations were addressed in the process of constructing the curated datasets. More than a mere collection of facts, the documentation seeks to organize these diverse set of information into a coherent story, allowing end-users of the curated datasets to more quickly (1) see to what extent is it feasible to investigate a particular scientific question using the data on-hand, and (2) see to what extent is it feasible to consider potential advancements in computational methodology.

GOAL: By the end of this module, you will be able to:

- Correctly interpret columns in the curated datasets.
- Correctly identify rows which need to be included/excluded when working on (1) and/or (2) described above.

1 What do the rows represent?

Each row of `merged.csv` represents a successfully launched EMA questionnaire during the study...

Figure 1: Rows and Columns in Toy Dataset

...and we can tell

- what kind of EMA questionnaire was launched, i.e., what kind of survey was 'pushed' by the smartphone to the participant (`assessment_type`)

- when the EMA questionnaire was launched (`delivered_unixts`)
- whether the participant responded *any* item in the EMA questionnaire (`with_any_response`)
- if the participant responded to *any* item in the EMA questionnaire, when they began providing their responses (`begin_unixts`)

We say 'successfully launched' to emphasize that there may be prolonged periods of time when no EMA questionnaires of any type were launched (e.g., due to software issue or when phone was switched off). Hence, during such periods of time, no EMA questionnaires will be displayed on the participant's smartphone. Correspondingly, when performing a visual inspection of the data, such periods of time may be represented by a prolonged time gap between two consecutive EMA questionnaires in the `merged.csv` file.

Additionally, we note that although it may appear unnecessary to have an indicator for whether the participant responded to Self-Initiated EMA Questionnaires (i.e., the variable `with_any_response` was constructed for all types of EMA Questionnaires), it is possible (but uncommon) for a participant to eventually decide to ignore Self-Initiated EMA Questionnaires (i.e., they perform a button press, and the button press was followed by the launching of an EMA Questionnaire, but for whatever reason, the participant decided to not follow through with responding to the EMA Questionnaire).

We also need a way to chronologically order EMAs. While completed and partially completed EMAs may have a timestamp for Begin Time (`begin_unixts`), EMAs which have no response will not have a timestamp for Begin Time. Even so, we may still be able to chronologically order EMAs if we consider timestamps for Delivered Time (`delivered_unixts`) as well. More specifically, to chronologically order EMAs within `merged.csv`, we may utilize the time variable Aligned Time (`time_unixts`) which is equal to Begin Time (`begin_unixts`) if the participant responded to the EMA (`with_any_response=1`) and equal to Delivered Time (`delivered_unixts`) if the participant did not respond to the EMA (`with_any_response=0`).

Let us place markers chronologically on a timeline. These markers represent Aligned Time for Random EMA (triangles) and Urge EMA (squares).

Figure 2: Visual Diagram for Toy Dataset: A Timeline

Next, let us draw an identical timeline directly on top to visualize how many among the EMAs successfully launched had any response to the EMA questionnaire (`with_any_response=1`).

Figure 3: Visual Diagram for Toy Dataset: A Timeline

2 What information are *not* represented in the rows of `merged.csv`?

Recall that out of the 200 participants enrolled in the PNS study, 37 participants will be excluded from all data analyses. Any responses provided by these 37 participants during the conduct of the study have been

excluded from `merged.csv`. Hence, none of the rows `merged.csv` will contain any information from these 37 participants. The IDs of these participants can be referenced in the file `quit_dates_final.csv`. In the `quit_dates_final.csv` file, these participants can be identified by the indicator variable `exclude`, which is equal to 1 if the ID belongs to one of these 37 participants, and equal to 0, if the ID dies not belong to one of these 37 participants.

3 What do the columns represent?

The columns suffixed by `_item_XX` refer to items within EMA questionnaires...

Figure 4: Rows and Columns in Toy Dataset

...and we can tell

- which EMA questionnaire the column is applicable to (e.g., `postquit_alreadyslipped_item_7` represents an item in the Post-Quit Already Slipped Questionnaire; `postquit_random_item_7` represents an item in the Post-Quit Random EMA Questionnaire)
- what were the original responses provided by participants to the questionnaires; values under the columns suffixed by `_item_XX` were left unchanged from the raw datasets

The curated smoking variables ...

- `smoking_indicator` - an indicator for whether *any* (1 = at least one; 0 = none) cigarette smoking occurred between the *last assessment* and the *current assessment* (i.e., the EMA Questionnaire from which the value of this variable was derived)
- `smoking_qty` - a count of the number of cigarettes smoked (0, 0.5, 1.5, 3.5, 5.5, 7.5, 9.5) between the *last assessment* and the *current assessment* (i.e., the EMA Questionnaire from which the value of this variable was derived)
- `smoking_delta_minutes` - the number of minutes prior to the *current assessment* (i.e., the EMA Questionnaire from which the value of this variable was derived) when the reported cigarettes smoked in `smoking_qty` occurred; more specifically, if more than one cigarette was smoked, this variable captures the time when that final cigarette stick just prior to the *current assessment* was smoked

... are provided side-by-side the original responses to EMA questionnaires.

Figure 5: Rows and Columns in Toy Dataset

Figure 6: Rows and Columns in Toy Dataset

BREAK: Any questions?

Note that not all rows in the `merged.csv` file may be used to construct a dependent variable on smoking behavior. The variable `ema_order` allows us to quickly identify rows which may be used: a missing value represents the fact that the row may not be used but a non-missing value represents the fact that the row may be used. More specifically we have that,

Case #	Type of EMA Questionnaire (`assessment_type`)	Did participant provide any response? (`with_any_response`)	May this row be used?	Value of `ema_order`
1	Pre-/Post- Quit Random EMA	They provided a response (1)	Yes	1, 2, 3, ...
2	Pre-/Post- Quit Random EMA	They did not provide a response (0)	No	missing
3	Pre-/Post- Quit Self-Initiated EMA	They provided a response (1)	Yes	1, 2, 3, ...
4	Pre-/Post- Quit Self-Initiated EMA	They did not provide a response (0)	Yes	1, 2, 3, ...

Recall that participants are asked to report, 'Since *the last assessment*, have you smoked any cigarettes that you did not record?' Hence, the effect of excluding Random EMAs for which the participant did not provide any response is to assume that when the participant recalls smoking information relative to the 'last assessment', that they are only recalling EMAs in Case # 1, 3, 4 above.

Let us reinforce the concept of what it means for a participant to have provided information in a particular EMA Questionnaire *relative to the last assessment* through visual diagrams.

Figure 7: Rows and Columns in Toy Dataset

Figure 8: Rows and Columns in Toy Dataset

BREAK: Any questions?

Let's see how these concepts are used as we break apart `merged.csv` into bite-sized chunks of data files which may be used as a starting point for data analysis. The emphasis of this section (and others like it) will be on the workflow and logic of how one would work through grabbing rows/files, rather than `R` syntax.

4.1 Step 1

```
# Read in data file
dat_big_merged <- read.csv(file.path(path_pns_input_data, "merged.csv"),
                           header = TRUE,
                           na.strings = "")

# View the first few rows of the data file
head(dat_big_merged)
```

4.2 Step 2

```
# How do we get from merged.csv to post_quit_random_ema.csv?
dat_postquit_random_ema <- dat_big_merged %>%
  # Select columns corresponding to participant ID and time variables;
  # select columns corresponding to items in Post-Quit Random EMA questionnaire
  select(id:time_unixts, postquit_random_item_1:postquit_random_item_67) %>%
  # Select those rows corresponding to when Post-Quit Random EMA was launched
  filter(assessment_type == "Post-Quit Random") %>%
  # Remember: order according to increasing participant ID
  # and within each participant ID, according to increasing time
  arrange(id, time_unixts)

# How do we get from merged.csv to pre_quit_random_ema.csv?
dat_prequit_random_ema <- dat_big_merged %>%
  # Select columns corresponding to participant ID and time variables;
  # select columns corresponding to items in Pre-Quit Random EMA questionnaire
  select(id:time_unixts, prequit_random_item_1:prequit_random_item_67) %>%
  # Select those rows corresponding to when Pre-Quit Random EMA was launched
  filter(assessment_type == "Pre-Quit Random") %>%
  # Remember: order according to increasing participant ID
  # and within each participant ID, according to increasing time
  arrange(id, time_unixts)
```

Although we anticipate that the vast majority of papers will exclusively use responses from Pre-/Post- Quit Random EMA Questionnaires as Independent Variables (IV's), there may be situations where one may want to consider using responses from other kinds of EMA questionnaires. For example, responses from Pre-/Post- Quit Urge EMA Questionnaires when craving (rather than smoking) is the primary Dependent Variable (DV) of interest. In such situations, we may apply a similar logic as the code snippet above to obtain those rows and columns corresponding to Pre-/Post- Quit Urge EMA Questionnaires. An example code snippet applying the logic for Pre-/Post- Quit Urge EMA Questionnaires is displayed below.

```r
# How do we get from merged.csv to post_quit_urge_ema.csv?
dat_postquit_urge_ema <- dat_big_merged %>%
  # Select columns corresponding to participant ID and time variables;
  # select columns corresponding to items in Post-Quit Urge EMA questionnaire
  select(id:time_unixts, postquit_urge_item_1:postquit_urge_item_67) %>%
  # Select those rows corresponding to when Post-Quit Urge EMA was launched
  filter(assessment_type == "Post-Quit Urge") %>%
  # Remember: order according to increasing participant ID
  # and within each participant ID, according to increasing time
  arrange(id, time_unixts)

# How do we get from merged.csv to pre_quit_urge_ema.csv?
dat_prequit_urge_ema <- dat_big_merged %>%
  # Select columns corresponding to participant ID and time variables;
  # select columns corresponding to items in Pre-Quit Urge EMA questionnaire
  select(id:time_unixts, prequit_urge_item_1:prequit_urge_item_67) %>%
  # Select those rows corresponding to when Pre-Quit Urge EMA was launched
  filter(assessment_type == "Pre-Quit Urge") %>%
  # Remember: order according to increasing participant ID
  # and within each participant ID, according to increasing time
  arrange(id, time_unixts)
```

4.3 Step 3

```r
# How to we get from merged.csv to smoking.csv?
dat_smoking <- dat_big_merged %>%
  select(id:smoking_delta_minutes) %>%
  # We consider the "last assessment" to refer to either of the two situations below:
  # 1. participant-initiated EMAs (any type) having
  # with_any_response=0 or with_any_response=1
  # 2. Random EMA having with_any_response=1
  # In other words, the only situation not included in the "last assessment"
  # are those Random EMAs which the participant did not provide any response
  # All rows in merged.csv having
  filter(!is.na(ema_order)) %>%
  # Remember: order according to increasing participant ID
  # and within each participant ID, according to increasing time
  arrange(id, time_unixts)
```