01: Clean Data

2026-01-30

Contents

Setup **1**
- Load Libraries . . . 1
- Set File Paths . . . 1
- Load Processed Data . . . 1

Clean Data **1**
- Remove Irrelevant Columns . . . 2
- Apply Inclusion Criteria and Final Variable Selection . . . 2

Save and Verify **2**
- Save Cleaned Dataset . . . 2
- Quick Check . . . 2

Setup

Load Libraries

```
library(readxl)
library(dplyr)
```

Set File Paths

```
root <- "/Users/pei-chin/Research/Behavioral Science and Marketing_data_task"

data_raw       <- file.path(root, "data_raw")
data_processed <- file.path(root, "data_processed")
data_clean     <- file.path(root, "data_clean")
scripts        <- file.path(root, "scripts")
output         <- file.path(root, "output")
figures        <- file.path(output, "figures")
```

Load Processed Data

```
load(file.path(data_processed, "processed_data.RData"))
```

Clean Data

Sanity checks have been completed and documented in the script `00_Appendix: Data Quality Checks`. This section focuses on dropping variables unnecessary for analysis.

Remove Irrelevant Columns

These column indices correspond to metadata, system-generated fields, and other variables that are not relevant for analysis, so they are removed here.

```
df_data <- processed_data %>%
  select(-c(1:6), -c(8:17), -c(19:26), -c(62, 64, 65, 68, 69, 70))
```

Apply Inclusion Criteria and Final Variable Selection

```
clean_data <- df_data %>%
  # Apply inclusion criteria
  filter(Finished == "TRUE", consent == "AGREE", passedattn == "yes") %>%
  # Remove variables not needed for analysis
  # Some variables were not shown to participants in the survey interface,
  # so they are removed here to avoid ambiguity in interpretation
  select(
    -real_imaginary, -initials_box, -describe, -real_imaginary, -describe,
    -feelings_exp, -attention_1, -attention_1_TEXT, -attention_2, -attention_2_TEXT,
    -attention_3, -attention_3_TEXT, -target_sex_3_TEXT, -sex_3_TEXT, -comments,
    -consent, -Finished, -passedattn, -initials, -initiator_type
  ) %>%
  # Create participant ID

  mutate(id = row_number()) %>%
  # Create a high vs. low self-blame grouping variable using a median split at 50
  # This step distinguishes participants by their tendency toward self-blame
  # While dichotomizing a continuous variable has methodological limitations,
  # a binary grouping (0/1) is used here for simplicity and time constraints
  mutate(high_blame = if_else(blame_1 > 50, 1, 0)) %>%
  # Move ID to the first column for easier inspection and reference
  relocate(id)
```

Save and Verify

Save Cleaned Dataset

```
save(
  clean_data,
  file = file.path(data_clean, "clean_data.RData")
)
```

Quick Check

```
dim(clean_data)
```

```
## [1] 45 24
```