Language (Covid CNS)

Abigail ter Kuile

31/01/2022

Set up

Delete everything in your global environment

```
remove(list = ls())
```

Add the add_numeric function - used to convert character variables into numeric variables. Add the sumscores function - used to generate sumscores Add the package_check function - used to install and load dependencies

```
source(file = "../functions/add_numeric_1.R")
source(file = "../functions/remove_duplicates.R")
source(file = "../functions/sumscores.R")
source(file = "../functions/package_check.R")
source(file = "../functions/imp_check.R")
```

Note: always load tidyverse last

```
packages = c(
  "summarytools",
  "sjlabelled",
  "Amelia",
  "gtsummary",
  "tidyverse"
  )
package_check(packages)
```

```
Loading required package: summarytools

Warning: package 'summarytools' was built under R version 4.0.5

Registered S3 method overwritten by 'pryr':
  method      from
  print.bytes Rcpp

Loading required package: sjlabelled
```

```
Attaching package: 'sjlabelled'

The following object is masked from 'package:summarytools':

    unlabel

Loading required package: Amelia

Warning: package 'Amelia' was built under R version 4.0.5

Loading required package: Rcpp

Warning: package 'Rcpp' was built under R version 4.0.5

##
## Amelia II: Multiple Imputation
## (Version 1.8.0, built: 2021-05-26)
## Copyright (C) 2005-2022 James Honaker, Gary King and Matthew Blackwell
## Refer to http://gking.harvard.edu/amelia/ for more information
##

Loading required package: gtsummary

Warning: package 'gtsummary' was built under R version 4.0.5

Loading required package: tidyverse

Warning: package 'tidyverse' was built under R version 4.0.5

-- Attaching packages --------------------------------------- tidyverse 1.3.1 --

v ggplot2 3.3.5     v purrr   0.3.4
v tibble  3.1.5     v dplyr   1.0.7
v tidyr   1.1.4     v stringr 1.4.0
v readr   2.0.2     v forcats 0.5.1

Warning: package 'ggplot2' was built under R version 4.0.5

Warning: package 'tibble' was built under R version 4.0.5

Warning: package 'tidyr' was built under R version 4.0.5

Warning: package 'readr' was built under R version 4.0.5

Warning: package 'purrr' was built under R version 4.0.5

Warning: package 'dplyr' was built under R version 4.0.5

Warning: package 'stringr' was built under R version 4.0.5
```

```
Warning: package 'forcats' was built under R version 4.0.5
```

```
-- Conflicts ----------------------------------------- tidyverse_conflicts() --
x forcats::as_factor() masks sjlabelled::as_factor()
x dplyr::as_label()    masks ggplot2::as_label(), sjlabelled::as_label()
x dplyr::filter()      masks stats::filter()
x dplyr::lag()         masks stats::lag()
x tibble::view()       masks summarytools::view()
```

Retrieve the recent date

We are using the recent date to save files with paste0() as an extension to not overwrite old versions

```
date = Sys.Date()
date
```

```
[1] "2022-02-22"
```

Read in file with path to ilovedata channel on Teams

```
source(file = "../credentials/paths.R")
```

Read in the data: Demographics

COVID CNS data

```
covidcns_dat <- read_rds(
  file = paste0(ilovedata, "/data_raw/latest_freeze/covid_cns/baseline/dem_covid_cns.rds")
  )

# check variable names in dataframe
covidcns_dat %>%
  colnames()
```

```
  [1] "externalDataReference"
  [2] "startDate"
  [3] "endDate"
  [4] "dem.day"
  [5] "dem.month"
  [6] "dem.year"
  [7] "dem.required_question_eligibility_criteria.txt"
  [8] "dem.what_gender_do_you_identify_with"
  [9] "dem.what_gender_do_you_identify_with.txt"
 [10] "dem.do_you_consider_yourself_to_be_transgender"
 [11] "dem.have_you_ever_been_pregnant"
 [12] "dem.what_is_your_sexual_orientation"
 [13] "dem.what_is_your_sexual_orientation.txt"
```

```
[14] "dem.what_is_your_current_maritalrelationship_status"
[15] "dem.what_is_your_current_maritalrelationship_status.txt"
[16] "dem.how_would_you_describe_your_vision"
[17] "dem.how_would_you_describe_your_hearing"
[18] "dem.which_hand_do_you_usually_write_with"
[19] "dem.college_or_university_degree"
[20] "dem.a_levelsas_levels_or_equivalent"
[21] "dem.o_levelsgcses_or_equivalent"
[22] "dem.cses_or_equivalent"
[23] "dem.nvq_or_hnd_or_hnc_or_equivalent"
[24] "dem.other_professional_qualifications_"
[25] "dem.other_professional_qualifications_text.txt"
[26] "dem.none_of_the_above"
[27] "dem.prefer_not_to_say"
[28] "dem.british_mixed_british"
[29] "dem.irish"
[30] "dem.northern_irish"
[31] "dem.any_other_white_background"
[32] "dem.white_and_black_caribbean"
[33] "dem.white_and_black_africa"
[34] "dem.white_and_asian"
[35] "dem.any_other_mixed_background"
[36] "dem.indian_or_british_indian"
[37] "dem.pakistani_or_british_pakistani"
[38] "dem.bangladeshi_or_british_bangladeshi"
[39] "dem.any_other_asian_background"
[40] "dem.caribbean"
[41] "dem.african"
[42] "dem.any_other_black_background"
[43] "dem.chinese"
[44] "dem.any_other_ethnic_group"
[45] "dem.other"
[46] "dem.othertext.txt"
[47] "dem.english"
[48] "dem.scottish"
[49] "dem.welsh"
[50] "dem.cornish"
[51] "dem.cypriot_"
[52] "dem.greek"
[53] "dem.greek_cypriot"
[54] "dem.italian"
[55] "dem.irish_traveller"
[56] "dem.traveller"
[57] "dem.gypsyromany"
[58] "dem.polish"
[59] "dem.republics_made_ussr"
[60] "dem.kosovan"
[61] "dem.albanian"
[62] "dem.bosnian"
[63] "dem.croatian"
[64] "dem.serbian"
[65] "dem.republics_made_yugoslavia"
[66] "dem.mixed_white"
[67] "dem.other_white_european_european_unspecified_european_mix"
```

```
 [68] "dem.black_and_asian"
 [69] "dem.black_and_chinese"
 [70] "dem.black_and_white"
 [71] "dem.chinese_and_white"
 [72] "dem.asian_and_chinese"
 [73] "dem.other_mixed_mixed_unspecified"
 [74] "dem.other_mixed_mixed_unspecifiedtext.txt"
 [75] "dem.mixed_asian"
 [76] "dem.punjabi"
 [77] "dem.kashmiri"
 [78] "dem.east_african_asian"
 [79] "dem.tamil"
 [80] "dem.sinhalese"
 [81] "dem.british_asian"
 [82] "dem.caribbean_asian"
 [83] "dem.other_asian_asian_unspecified"
 [84] "dem.other_asian_asian_unspecifiedtext.txt"
 [85] "dem.somali"
 [86] "dem.mixed_black"
 [87] "dem.nigerian"
 [88] "dem.black_british"
 [89] "dem.other_black_black_unspecified"
 [90] "dem.other_black_black_unspecifiedtext.txt"
 [91] "dem.is_english_your_first_language"
 [92] "dem.what_is_your_first_language"
 [93] "dem.what_is_your_first_language.txt"
 [94] "dem.please_select_your_preferred_units_of_measurement"
 [95] "dem.what_is_your_current_height"
 [96] "dem.what_is_your_current_height.1"
 [97] "dem.what_is_your_current_height.2"
 [98] "dem.pregnant_weigh_weight_provide"
 [99] "dem.pregnant_weigh_weight_provide.1"
[100] "dem.pregnant_weigh_weight_provide.2"
[101] "dem.pregnant_weighed_weight_provide"
[102] "dem.pregnant_weighed_weight_provide.1"
[103] "dem.pregnant_weighed_weight_provide.2"
[104] "dem.highest_weight"
[105] "dem.stopped_growing_adult_height"
[106] "dem.stopped_growing_adult_height.1"
[107] "dem.stopped_growing_adult_height.2"
[108] "dem.body_suffered_injury_involving"
[109] "dem.middle_wake_night_covid19"
[110] "dem.middle_wake_night_covid19.1"
[111] "dem.medical_history_birth_relevant"
[112] "dem.affects_concerned_live_memory"
[113] "dem.memory_problem_worse_year"
[114] "dem.based_confirm_living_question"
[115] "dem.diagnosed_required_question_covid19"
[116] "dem.long_ago_diagnosed_required"
[117] "dem.long_ago_diagnosed_required.1"
[118] "dem.diagnosed_covid19_experienced_similar"
[119] "dem.quality_rate_life"
[120] "dem.energy_everyday_life"
[121] "dem.opportunity_leisure_activities"
```

```
[122] "dem.money_day"
[123] "dem.middle_wake_night_trouble"
[124] "dem.affects_concerned_live_memory.1"
[125] "dem.affects_concerned_live_memory.2"
[126] "dem.has_your_memory_got_progressively_worse"
[127] "dem.vietnamese"
[128] "dem.filipino"
[129] "dem.malaysian"
[130] "dem.any_other_group"
[131] "dem.any_other_grouptext.txt"
[132] "dem.lowest_weight_adult_height"
[133] "dem.happy_general_health"
```

```
# Inspect dimensions of dataframe (number of rows and columns)
covidcns_dat %>%
  dim()
```

```
[1] 235 133
```

Specify columns to be excluded from add_numeric function Continuous variables should be excluded, as they are already numeric NB: If this is data from the COPING survey, add "_cop" to the end of each variable name

```
exclude_cols_numeric <- c(
  "ID",
  "sample",
  "startDate",
  "endDate",
  "dem.what_is_your_first_language.txt"
  )
```

Select & rename relevant columns (will be a function at some point)

```
covidcns_dat_id <- covidcns_dat %>% # new dataset with ID
  drop_na(externalDataReference) %>% # Drop participants with no ID
  distinct(externalDataReference, .keep_all = TRUE) %>% # Changed to distinct due to NA coercion
  add_column(sample = "COVIDCNS",
             .after = "externalDataReference") %>% # Create new sample column
  select(
         ID = externalDataReference,# ID
         sample,
         startDate,
         endDate,
         dem.what_is_your_first_language,
         dem.what_is_your_first_language.txt,
         dem.is_english_your_first_language
         )  %>%

 add_numeric_1(exclude = exclude_cols_numeric)

# Inspect colnames
covidcns_dat_id %>%
  colnames()
```

```
[1] "ID"
[2] "sample"
[3] "startDate"
[4] "endDate"
[5] "dem.what_is_your_first_language"
[6] "dem.is_english_your_first_language"
[7] "dem.what_is_your_first_language.txt"
[8] "dem.what_is_your_first_language_numeric"
[9] "dem.is_english_your_first_language_numeric"
```

Look at number of people excluded

```
# Inspect dimensions of new data set
covidcns_dat_id %>%
  dim()
```

```
[1] 228   9
```

```
# Inspect number of rows dropped
covidcns_excluded <- dim(covidcns_dat_id)[1]-dim(covidcns_dat)[1]
covidcns_excluded
```

```
[1] -7
```

Inspect numeric variables

```
covidcns_dat_id %>%
  select(all_of(ends_with("numeric"))) %>%
  tbl_summary(missing_text = "Missing")
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 228**
What is your first language?	20 (15, 24)
Missing	196
Is English your first language?	195 (86%)
Missing	1

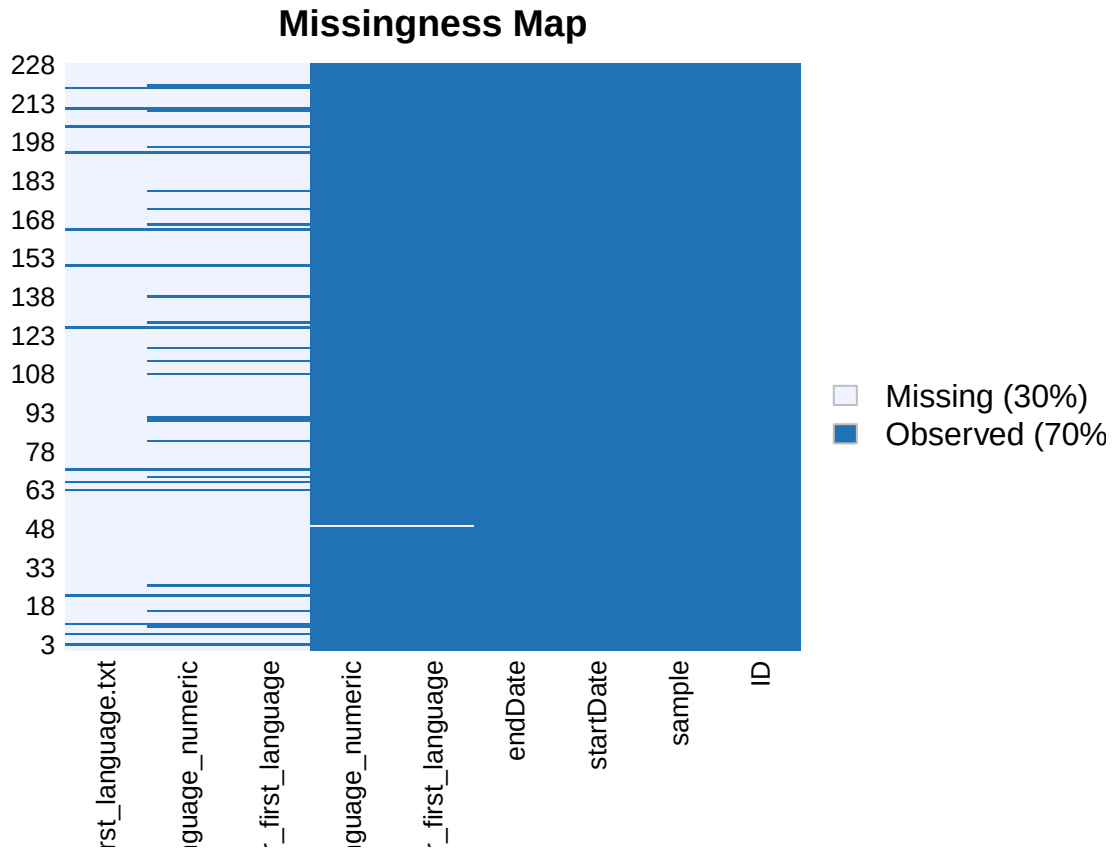
Check missingness by missmap

```
 covidcns_miss_map <- covidcns_dat_id %>%
   missmap()
```

```
Warning: Unknown or uninitialised column: `arguments`.

Warning: Unknown or uninitialised column: `arguments`.
```

```
Warning: Unknown or uninitialised column: 'imputations'.
```




```
covidcns_miss_map
```

```
NULL
```

Rename covidcns_dat_id to dat

```
dat <- covidcns_dat_id
```

Data cleaning

Recode Non-answer values to 3 digits -555 'Not applicable' response from participant -777 Seen but not answered -888 Don't know -999 Prefer not to answer/Prefer not to say `NA` Were not shown the question (genuinely missing value) When we code someone as being 'not applicable' by deduction, we use `NA_real_`

```
dat <- dat %>%
  mutate(across(ends_with("numeric"),
                ~case_when(
                  . == -55 ~ -555,
                  . == -77 ~ -777,
                  . == -88 ~ -888,
                  . == -99 ~ -999,
                  TRUE ~ .)))
```

Cleaning numeric variables

Numeric 1-24 variables

Note: English as first language is asked as a separate question. dem.what_is_your_first_language is coded 1-4 and 6-24 (5 is missing - which should be English) Recode English as 5 under dem.what_is_your_first_language

```
dat <- dat %>%
  mutate(
    dem.what_is_your_first_language_numeric =
      if_else(
        dem.is_english_your_first_language_numeric == 1,
        true = 5,
        false = dem.what_is_your_first_language_numeric,
        missing = NA_real_
      )
  )

#check
dat %>%
  select(dem.is_english_your_first_language_numeric,
         dem.what_is_your_first_language_numeric)
```

```
# A tibble: 228 x 2
   dem.is_english_your_first_language_numeric dem.what_is_your_first_language_n~
                                        <dbl>                               <dbl>
 1                                          1                                   5
 2                                          1                                   5
 3                                          1                                   5
 4                                          1                                   5
 5                                          1                                   5
 6                                          1                                   5
 7                                          1                                   5
 8                                          1                                   5
 9                                          0                                  19
10                                          0                                  24
# ... with 218 more rows
```

Cleaning numeric variables

```
values_numeric_24 <- c(
  1,
  2,
  3,
  4,
  5,
  6,
  7,
  8,
  9,
  10,
  11,
```

```
  12,
  13,
  14,
  15,
  16,
  17,
  18,
  19,
  20,
  21,
  22,
  23,
  24,
  -777,
  NA
  )
values_numeric_24
```

```
 [1]    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
[16]   16   17   18   19   20   21   22   23   24 -777   NA
```

Create vector of variable names for numeric variables

```
variables_numeric_24 <- c(
    "dem.what_is_your_first_language_numeric"
  )
variables_numeric_24
```

```
[1] "dem.what_is_your_first_language_numeric"
```

Use imp_check function to find if any implausible values and obtain summary table of variables

```
imp_check(data = dat,
          variables = variables_numeric_24,
          values = values_numeric_24)
```

```
[1] "There are no implausible values in the dataset. Can leave these variables as they are."

Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 228**
dem.what_is_your_first_language_numeric	5.0 (5.0, 5.0)
Missing	1

Numeric 0-1 variables

Cleaning numeric variables

```
values_numeric_0 <- c(
  0,
  1,
  -999,
  NA
  )
values_numeric_0
```

```
[1]    0    1 -999   NA
```

Create vector of variable names for numeric variables

```
variables_numeric_0 <- c(
    "dem.is_english_your_first_language_numeric"
  )
variables_numeric_0
```

```
[1] "dem.is_english_your_first_language_numeric"
```

Use imp_check function to find if any implausible values and obtain summary table of variables

```
imp_check(data = dat,
           variables = variables_numeric_0,
          values = values_numeric_0)
```

```
[1] "There are no implausible values in the dataset. Can leave these variables as they are."
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 228**
dem.is_english_your_first_language_numeric	195 (86%)
Missing	1

Cleaning Categorical variables

Categorical variables 1-24

Add categorical label "English" to dem.what_is_your_first_language

```
dat <- dat %>%
  mutate(
    dem.what_is_your_first_language =
      case_when(
        dem.what_is_your_first_language_numeric == 5 ~ "English",
```

```
        TRUE ~ as.character(dem.what_is_your_first_language)
      )
  )
#check
dat %>%
  select(dem.what_is_your_first_language,
         dem.what_is_your_first_language_numeric)
```

```
# A tibble: 228 x 2
   dem.what_is_your_first_language dem.what_is_your_first_language_numeric
   <chr>                                                             <dbl>
 1 English                                                               5
 2 English                                                               5
 3 English                                                               5
 4 English                                                               5
 5 English                                                               5
 6 English                                                               5
 7 English                                                               5
 8 English                                                               5
 9 Spanish                                                              19
10 Other                                                                24
# ... with 218 more rows
```

Create vector of categorical values for variables

```
values_categorical_24 <- c(
  "Arabic",
  "Bengali",
  "Chinese",
  "Danish",
  "English",
  "French",
  "Gaelic",
  "German",
  "Hindi",
  "Japanese",
  "Javanese",
  "Korean",
  "Lahnda",
  "Mandarin",
  "Polish",
  "Portuguese",
  "Punjabi",
  "Russian",
  "Spanish",
  "Tamil",
  "Turkish",
  "Vietnamese",
  "Welsh",
  "Other",
  "Seen but not answered",
  NA
```

```
  )
values_categorical_24
```

```
 [1] "Arabic"                "Bengali"               "Chinese"
 [4] "Danish"                "English"               "French"
 [7] "Gaelic"                "German"                "Hindi"
[10] "Japanese"              "Javanese"              "Korean"
[13] "Lahnda"                "Mandarin"              "Polish"
[16] "Portuguese"            "Punjabi"               "Russian"
[19] "Spanish"               "Tamil"                 "Turkish"
[22] "Vietnamese"            "Welsh"                 "Other"
[25] "Seen but not answered" NA
```

Create vector of variable names for categorical variables

```
variables_categorical_24 <-
  c(
    "dem.what_is_your_first_language"
    )
variables_categorical_24
```

```
[1] "dem.what_is_your_first_language"
```

Use imp_check function to find if any implausible values and obtain summary table of variables

```
imp_check(data = dat,
          variables = variables_categorical_24,
          values = values_categorical_24)
```

```
[1] "There are no implausible values in the dataset. Can leave these variables as they are."
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	**N = 228**
dem.what_is_your_first_language	
Arabic	2 (0.9%)
Bengali	1 (0.4%)
Chinese	1 (0.4%)
English	195 (86%)
French	1 (0.4%)
Hindi	1 (0.4%)
Other	14 (6.2%)
Polish	4 (1.8%)
Punjabi	1 (0.4%)
Russian	1 (0.4%)
Spanish	4 (1.8%)
Turkish	1 (0.4%)

Characteristic	N = 228
Welsh	1 (0.4%)
Missing	1

Categorical variables 0-1

Create vector of categorical values for variables

```
values_categorical_0 <- c(
  "No",
  "Yes",
  "Prefer not to say",
  NA)
values_categorical_0
```

```
[1] "No"                "Yes"               "Prefer not to say"
[4] NA
```

Create vector of variable names for categorical variables

```
variables_categorical_0 <-
  c(
    "dem.is_english_your_first_language"
    )
variables_categorical_0
```

```
[1] "dem.is_english_your_first_language"
```

Use imp_check function to find if any implausible values and obtain summary table of variables

```
imp_check(data = dat,
          variables = variables_categorical_0,
          values = values_categorical_0)
```

```
[1] "There are no implausible values in the dataset. Can leave these variables as they are."
```

```
Table printed with `knitr::kable()`, not {gt}. Learn why at
http://www.danieldsjoberg.com/gtsummary/articles/rmarkdown.html
To suppress this message, include `message = FALSE` in code chunk header.
```

Characteristic	N = 228
Is English your first language?	
Prefer not to say	0 (0%)
Seen but not answered	0 (0%)
No	32 (14%)
Yes	195 (86%)

Characteristic	N = 228
Missing	1

Save cleaned data

Check colnames before exporting final dataset

```
colnames(dat)
```

```
[1] "ID"
[2] "sample"
[3] "startDate"
[4] "endDate"
[5] "dem.what_is_your_first_language"
[6] "dem.is_english_your_first_language"
[7] "dem.what_is_your_first_language.txt"
[8] "dem.what_is_your_first_language_numeric"
[9] "dem.is_english_your_first_language_numeric"
```

```
dat %>%
  filter(sample == "COVIDCNS") %>%  # select only COVID CNS participants
  saveRDS(
    file = paste0(ilovedata,
                  "/data/latest_freeze/covidcns/language_covidcns_clean.rds")
    )
```