format_check_cell_suppression

Claire Boulange

2023-05-10

```
# Load any necessary packages
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)
library(readr)

# Set any global options
options(digits = 3)
```

Open a connection to the output file Set the working directory - folder with data cleaned ready for QA

Define a list of acceptable values for the first column Define a function to check if column names are in snake case format Define a regex pattern to check age_group format Define regex patterns for age_group, year_range, and calendar_year

Get a list of CSV files in the working directory Print the number of files in the folder and their names

initialize an empty list to store data frame read each CSV file into a data frame and store it in the df_list append the data frame to the list using the file name as the key

Loop through each data frame in the list and print the head

```
for (df_name in names(df_list)) {
  cat(paste("Head of", df_name, ":\n"))
  print(head(df_list[[df_name]]))
}
```

```
## Head of aihw_3121_child_protection_substantiations_national.csv :
##   STE_CODE16 sex year_range age_group n_care_and_protection_orders
## 1          0 all  2016-2017      0-24                        12299
## 2          0 all  2017-2018      0-24                        12020
## 3          0 all  2018-2019      0-24                        12911
## 4          0 all  2019-2020      0-24                        13062
## 5          0 all  2020-2021      0-24                        12118
## Head of aihw_3121_child_protection_substantiations_STE.csv :
##   STE_CODE16 sex year_range age_group n_care_and_protection_orders
## 1          1 all  2016-2017      0-24                         2822
## 2          2 all  2016-2017      0-24                         4832
## 3          3 all  2016-2017      0-24                         1933
## 4          5 all  2016-2017      0-24                         1169
## 5          4 all  2016-2017      0-24                          746
## 6          6 all  2016-2017      0-24                          298
## Head of aihw_361_children_admitted_out_of_home_care_national.csv :
##   STE_CODE16 sex year_range age_group n_children_admitted_out_of_home_care
## 1          0 all  2016-2017       0-1                                 2245
## 2          0 all  2016-2017       1-4                                 3018
## 3          0 all  2016-2017       5-9                                 2795
## 4          0 all  2016-2017     10-14                                 2469
## 5          0 all  2016-2017     15-17                                 1119
## 6          0 all  2017-2018       0-1                                 2179
##   per_1000_children_admitted_out_of_home_care
## 1                                        7.15
## 2                                        2.39
## 3                                        1.77
## 4                                        1.70
## 5                                        1.29
## 6                                        7.14
## Head of aihw_361_children_admitted_out_of_home_care_STE.csv :
##   STE_CODE16 sex year_range age_group n_children_admitted_out_of_home_care
## 1          1 all  2016-2017       0-1                                  657
## 2          2 all  2016-2017       0-1                                  595
## 3          3 all  2016-2017       0-1                                  456
## 4          5 all  2016-2017       0-1                                  224
## 5          4 all  2016-2017       0-1                                  185
## 6          6 all  2016-2017       0-1                                   44
##   per_1000_children_admitted_out_of_home_care
## 1                                        6.54
## 2                                        7.36
## 3                                        7.39
## 4                                        6.35
## 5                                        9.19
## 6                                        7.51
## Head of aihw_361_children_discharged_from_out_of_home_care__national.csv :
##   STE_CODE16 sex year_range age_group n_children_discharged_from_home_care
## 1          0 all  2016-2017       0-1                                  552
## 2          0 all  2016-2017       1-4                                 2212
## 3          0 all  2016-2017       5-9                                 2295
## 4          0 all  2016-2017     10-14                                 2168
## 5          0 all  2016-2017     15-17                                 3108
## 6          0 all  2017-2018       0-1                                  600
##   per_1000_children_discharged_from_home_care
```

```
## 1                                       1.76
## 2                                       1.75
## 3                                       1.46
## 4                                       1.49
## 5                                       3.58
## 6                                       1.97
## Head of aihw_361_children_discharged_from_out_of_home_care_STE.csv :
##   STE_CODE16 sex year_range age_group n_children_discharged_from_home_care
## 1          1 all  2016-2017       0-1                                  127
## 2          2 all  2016-2017       0-1                                  197
## 3          3 all  2016-2017       0-1                                  126
## 4          5 all  2016-2017       0-1                                   43
## 5          4 all  2016-2017       0-1                                   25
## 6          6 all  2016-2017       0-1                                    8
##   per_1000_children_discharged_from_home_care
## 1                                       1.26
## 2                                       2.44
## 3                                       2.04
## 4                                       1.22
## 5                                       1.24
## 6                                       1.36
```

print overview of all CSV files and their variables:

```
csv_info <- data.frame(csv_file = names(df_list), variables = sapply(df_list, function(x) paste(names(x)
print(csv_info)
```

```
##
## aihw_3121_child_protection_substantiations_national.csv                  aihw_3121_child_protection_
## aihw_3121_child_protection_substantiations_STE.csv                            aihw_3121_child_prote
## aihw_361_children_admitted_out_of_home_care_national.csv                aihw_361_children_admitted_
## aihw_361_children_admitted_out_of_home_care_STE.csv                         aihw_361_children_admi
## aihw_361_children_discharged_from_out_of_home_care__national.csv aihw_361_children_discharged_from_ou
## aihw_361_children_discharged_from_out_of_home_care_STE.csv             aihw_361_children_discharged_
##
## aihw_3121_child_protection_substantiations_national.csv
## aihw_3121_child_protection_substantiations_STE.csv
## aihw_361_children_admitted_out_of_home_care_national.csv         STE_CODE16, sex, year_range, age_gr
## aihw_361_children_admitted_out_of_home_care_STE.csv              STE_CODE16, sex, year_range, age_gr
## aihw_361_children_discharged_from_out_of_home_care__national.csv STE_CODE16, sex, year_range, age_gr
## aihw_361_children_discharged_from_out_of_home_care_STE.csv       STE_CODE16, sex, year_range, age_gr
```

This code loops through each data frame in df_list and creates a data frame var_info with columns for variable name, class and range. The class column is populated using the class function and the range column is populated using the range function and then collapsed into a string using paste. Finally, the var_info data frame is printed for each data frame in df_list.

you might want to use it to fill the dictionary:

https://connectqutedu.sharepoint.com/:x:/r/teams/FOS_PRO_ANCHDA/Shared%20Documents/General/Metadata/data_dictionary.xlsx?d=w8708a1fa697f42899fd11956f7bc1ce6&csf=1&web=1&e=W3Iqp4

```r
for (df_name in names(df_list)) {
  cat(paste("Table for", df_name, ":\n"))

  # Get the data frame
  df <- df_list[[df_name]]

  # Create a data frame with variable name, class and range
  var_info <- data.frame(variable = names(df),
                         class = sapply(df, class),
                         range = sapply(df, function(x) paste(range(x), collapse = " - ")))

  # Print the variable information
  print(var_info)
}
```

```
## Table for aihw_3121_child_protection_substantiations_national.csv :
##                                                 variable     class
## STE_CODE16                                     STE_CODE16   integer
## sex                                                   sex character
## year_range                                     year_range character
## age_group                                       age_group character
## n_care_and_protection_orders n_care_and_protection_orders   integer
##                                                    range
## STE_CODE16                                       0 - 0
## sex                                          all - all
## year_range                       2016-2017 - 2020-2021
## age_group                                  0-24 - 0-24
## n_care_and_protection_orders         12020 - 13062
## Table for aihw_3121_child_protection_substantiations_STE.csv :
##                                                 variable     class
## STE_CODE16                                     STE_CODE16   integer
## sex                                                   sex character
## year_range                                     year_range character
## age_group                                       age_group character
## n_care_and_protection_orders n_care_and_protection_orders   integer
##                                                    range
## STE_CODE16                                       1 - 8
## sex                                          all - all
## year_range                       2016-2017 - 2020-2021
## age_group                                  0-24 - 0-24
## n_care_and_protection_orders            102 - 5496
## Table for aihw_361_children_admitted_out_of_home_care_national.csv :
##                                                                                 variable
## STE_CODE16                                                                     STE_CODE16
## sex                                                                                   sex
## year_range                                                                     year_range
## age_group                                                                       age_group
## n_children_admitted_out_of_home_care                n_children_admitted_out_of_home_care
## per_1000_children_admitted_out_of_home_care per_1000_children_admitted_out_of_home_care
##                                                    class                 range
## STE_CODE16                                       integer                 0 - 0
## sex                                            character             all - all
## year_range                                     character 2016-2017 - 2020-2021
```

```
## age_group                                     character             0-1 - 5-9
## n_children_admitted_out_of_home_care          integer           1119 - 3044
## per_1000_children_admitted_out_of_home_care   numeric               NA - NA
## Table for aihw_361_children_admitted_out_of_home_care_STE.csv :
##                                                                          variable
## STE_CODE16                                                            STE_CODE16
## sex                                                                          sex
## year_range                                                            year_range
## age_group                                                              age_group
## n_children_admitted_out_of_home_care              n_children_admitted_out_of_home_care
## per_1000_children_admitted_out_of_home_care per_1000_children_admitted_out_of_home_care
##                                                    class                 range
## STE_CODE16                                       integer                 1 - 8
## sex                                            character             all - all
## year_range                                     character 2016-2017 - 2020-2021
## age_group                                      character             0-1 - 5-9
## n_children_admitted_out_of_home_care             integer             12 - 1134
## per_1000_children_admitted_out_of_home_care      numeric               NA - NA
## Table for aihw_361_children_discharged_from_out_of_home_care__national.csv :
##                                                                          variable
## STE_CODE16                                                            STE_CODE16
## sex                                                                          sex
## year_range                                                            year_range
## age_group                                                              age_group
## n_children_discharged_from_home_care              n_children_discharged_from_home_care
## per_1000_children_discharged_from_home_care per_1000_children_discharged_from_home_care
##                                                    class                 range
## STE_CODE16                                       integer                 0 - 0
## sex                                            character             all - all
## year_range                                     character 2016-2017 - 2020-2021
## age_group                                      character             0-1 - 5-9
## n_children_discharged_from_home_care             integer            552 - 3900
## per_1000_children_discharged_from_home_care      numeric          1.29 - 4.39
## Table for aihw_361_children_discharged_from_out_of_home_care_STE.csv :
##                                                                          variable
## STE_CODE16                                                            STE_CODE16
## sex                                                                          sex
## year_range                                                            year_range
## age_group                                                              age_group
## n_children_discharged_from_home_care              n_children_discharged_from_home_care
## per_1000_children_discharged_from_home_care per_1000_children_discharged_from_home_care
##                                                    class                 range
## STE_CODE16                                       integer                 1 - 8
## sex                                            character             all - all
## year_range                                     character 2016-2017 - 2020-2021
## age_group                                      character             0-1 - 5-9
## n_children_discharged_from_home_care             integer              2 - 1293
## per_1000_children_discharged_from_home_care      numeric         0.36 - 11.15
```

Iterate through each data frame in df_list and perform checks:

1. check if age_group, sex, calendar year or year_range columns exist
2. check if age_group values are in the correct format
3. check if year_range values are in the correct format

4. check if calendar_year values are in the correct format
5. check if geography column is one of the acceptable values

```
for (df_name in names(df_list)) {
  df <- df_list[[df_name]]
  col_names <- names(df)
  first_col <- colnames(df)[1]


  if (!("age_group" %in% col_names)) {
    cat("Error: age_group column not found in", df_name, "\n")
  }
  if (!("sex" %in% col_names)) {
    cat("Error: sex column not found in", df_name, "\n")
  }
  if (!("calendar_year" %in% col_names) && !("year_range" %in% col_names)) {
    cat("Error: either calendar_year or year_range column must be present in", df_name, "\n")
  } else if (("calendar_year" %in% col_names) && ("year_range" %in% col_names)) {
    cat("Error: both calendar_year and year_range columns cannot be present in", df_name, "\n")
  } else if ("calendar_year" %in% col_names && !all(grepl("\\d{4}", df$calendar_year))) {
    cat("Error: calendar_year values in", df_name, "are not in the correct format (expected format: \\d
  } else if ("year_range" %in% col_names && !all(grepl("\\d{4}-\\d{4}", df$year_range))) {
    cat("Error: year_range values in", df_name, "are not in the correct format (expected format: \\d{4}
  }


  if ("age_group" %in% col_names) {
    age_group_values <- df$age_group
    if (!all(grepl(age_group_regex, age_group_values))) {
      cat("Error: age_group values in", df_name, "are not in the correct format (expected format: \\d-\
      cat("Invalid values:\n")
      invalid_age_group_values <- age_group_values[!grepl(age_group_regex, age_group_values)]
      cat(paste(unique(invalid_age_group_values), collapse=", "), "\n")
    }
  }

  if ("year_range" %in% col_names) {
    year_range_values <- df$year_range
    if (!all(grepl(year_range_regex, year_range_values))) {
      cat("Error: year_range values in", df_name, "are not in the correct format (expected format: \\d{
      cat("Invalid values:\n")
      invalid_year_range_values <- year_range_values[!grepl(year_range_regex, year_range_values)]
      cat(paste(unique(invalid_year_range_values), collapse=", "), "\n")
    }
  }

  if ("calendar_year" %in% col_names) {
    calendar_year_values <- df$calendar_year
    if (!all(grepl(calendar_year_regex, calendar_year_values))) {
      cat("Error: calendar_year values in", df_name, "are not in the correct format (expected format: \
      cat("Invalid values:\n")
      invalid_calendar_year_values <- calendar_year_values[!grepl(calendar_year_regex, calendar_year_va
      cat(paste(unique(invalid_calendar_year_values), collapse=", "), "\n")
    }
```

```
  }

  if (!is_snake_case(col_names[-1])) {
    cat("Error: column names in", df_name, "are not in snake case format (lowercase words separated by u
  }

  if (!(first_col %in% first_col_check)) {
    cat("Error: geography column is not one of the acceptable values (", paste(first_col_check, collapse
  }

}
```

```
## Error: age_group values in aihw_361_children_admitted_out_of_home_care_national.csv are not in the co
## Invalid values:
## 1-4, 5-9, 10-14, 15-17
## Error: column names in aihw_361_children_admitted_out_of_home_care_national.csv are not in snake case
## Error: age_group values in aihw_361_children_admitted_out_of_home_care_STE.csv are not in the correct
## Invalid values:
## 1-4, 5-9, 10-14, 15-17
## Error: column names in aihw_361_children_admitted_out_of_home_care_STE.csv are not in snake case form
## Error: age_group values in aihw_361_children_discharged_from_out_of_home_care__national.csv are not i
## Invalid values:
## 1-4, 5-9, 10-14, 15-17
## Error: column names in aihw_361_children_discharged_from_out_of_home_care__national.csv are not in sn
## Error: age_group values in aihw_361_children_discharged_from_out_of_home_care_STE.csv are not in the
## Invalid values:
## 1-4, 5-9, 10-14, 15-17
## Error: column names in aihw_361_children_discharged_from_out_of_home_care_STE.csv are not in snake ca
```

CHECK OUTPUTS NOW!!!

Did you detect any formatting errors?

YES »> GO BACK TO YOUR CODE AND MAKE CORRECTIONS

NO »> PROCEED WITH CELL SUPPRESSION

Define the input directory path Define the output directory path to save cleaned dataset WITH CELL SUPPRESSED

```
input_dir <- "C:/Users/00095998/OneDrive - The University of Western Australia/acwa_temp/aihw_child_prot
output_dir <- "C:/Users/00095998/OneDrive - The University of Western Australia/acwa_temp/aihw_child_pro
```

Create the output directory if it doesn't already exist Get a list of all CSV files in the input directory

```
dir.create(output_dir, showWarnings = FALSE)
csv_files <- list.files(input_dir, pattern = ".csv$", full.names = TRUE)
```

Loop through each CSV file Check if there is an "uncertainty" column in the data frame

```
for (file in csv_files) {


  df <- read.csv(file, stringsAsFactors = FALSE)
```

```
  if ("uncertainty" %in% colnames(df)) {
    message(paste0("Note: The file ", basename(file), " contains an 'uncertainty' column. Make sure to r
  else {
  # Print message indicating that there is no need to apply cell suppression to "uncertainty" column
  cat("You don't have to worry about cell suppression on 'uncertainty' in", basename(file), "\n")
}
}
```

```
## You don't have to worry about cell suppression on 'uncertainty' in aihw_3121_child_protection_substar
## You don't have to worry about cell suppression on 'uncertainty' in aihw_3121_child_protection_substar
## You don't have to worry about cell suppression on 'uncertainty' in aihw_361_children_admitted_out_of_
## You don't have to worry about cell suppression on 'uncertainty' in aihw_361_children_admitted_out_of_
## You don't have to worry about cell suppression on 'uncertainty' in aihw_361_children_discharged_from_
## You don't have to worry about cell suppression on 'uncertainty' in aihw_361_children_discharged_from_
```

detect columns that are numeric and where you might need to apply cell suppression

```
# Define the exclusion list
exclude_list <- c("STE_CODE16", "SA2_CODE16", "SA3_CODE16", "SA4_CODE16", "LGA_CODE16", "Australia", "s

# Loop through each CSV file and check for columns that are numeric and not in the exclusion list
for (file in csv_files) {

  # Read in the CSV file
  df <- read.csv(file, stringsAsFactors = FALSE)

  # Get the names of columns that are numeric and not in the exclusion list
  num_cols <- names(df)[sapply(df, is.numeric) & !names(df) %in% exclude_list]

  # If there are any such columns, print a message for each file and column
  if (length(num_cols) > 0) {
    for (col in num_cols) {
      message(paste0("For file ", basename(file), ", check values in column '", col, "' for cell suppres
    }
  } else {
    # Print message indicating that there are no columns to check
    cat("You don't have to worry about cell suppression in any numeric columns in", basename(file), "\n
  }

}
```

```
## For file aihw_3121_child_protection_substantiations_national.csv, check values in column 'n_care_and_

## For file aihw_3121_child_protection_substantiations_STE.csv, check values in column 'n_care_and_prote

## For file aihw_361_children_admitted_out_of_home_care_national.csv, check values in column 'n_children

## For file aihw_361_children_admitted_out_of_home_care_national.csv, check values in column 'per_1000_

## For file aihw_361_children_admitted_out_of_home_care_STE.csv, check values in column 'n_children_admi
```

```
## For file aihw_361_children_admitted_out_of_home_care_STE.csv, check values in column 'per_1000_childr

## For file aihw_361_children_discharged_from_out_of_home_care__national.csv, check values in column 'n_

## For file aihw_361_children_discharged_from_out_of_home_care__national.csv, check values in column 'pe

## For file aihw_361_children_discharged_from_out_of_home_care_STE.csv, check values in column 'n_childr

## For file aihw_361_children_discharged_from_out_of_home_care_STE.csv, check values in column 'per_1000
```

Loop through each CSV file and apply cell suppression

```r
for (file in csv_files) {

  # Read in the CSV file
  df <- read.csv(file, stringsAsFactors = FALSE)

  # Apply cell suppression to 'n' column if it exists
  if ('n' %in% colnames(df)){
    df[!is.na(df[, "n"]) & df[, "n"] < 5, -(1:5)] <- 9999999
  }

  # Apply cell suppression to 'uncertainty' column
  if ('uncertainty' %in% colnames(df)){
    df[df[, "uncertainty"] == 2 & !is.na(df[, "uncertainty"]), colnames(df) %in% 'uncertainty' | colname
  }

  # Remove the 'uncertainty' column from the data frame
  df <- df[, !(colnames(df) %in% 'uncertainty')]

  # Write the modified data frame to a new CSV file in the output directory
  write.csv(df, file.path(output_dir, basename(file)), row.names = FALSE)

}
```