format_check_cell_suppression

Claire Boulange

2023-05-11

```
# 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)
```

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

```
cat("Number of files in folder:", length(csv_files), "\n")
```

```
## Number of files in folder: 4
```

```
for (file in csv_files) {
  if (!grepl("\\d{3,}", file)) {
    cat("WARNING! File(s) without indicator code:", file, "\n")
  }
}
cat(paste("File with indicator code:",csv_files, collapse = "\n"), "\n")
```

```
## File with indicator code: BITRE_172_children_0_16_motor_vehicle_accidents_STE.csv
## File with indicator code: BITRE_172_children_0_16_rolling_sum_over_10_years_motor_vehicle_accidents_S
## File with indicator code: BITRE_172_young_people_17_25_motor_vehicle_accidents_STE.csv
## File with indicator code: BITRE_172_young_people_17_25_rolling_sum_over_10_years_motor_vehicle_accide
```

```
cat("Finished checking CSV files.", "\n")
```

```
## Finished checking CSV files.
```

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 BITRE_172_children_0_16_motor_vehicle_accidents_STE.csv :
##   STE_CODE16    sex age_group calendar_year n_fatally_injured_in_road_accident
## 1          1 female      0-16          2008                                  3
## 2          1 female      0-16          2009                                 14
## 3          1 female      0-16          2010                                  6
## 4          1 female      0-16          2011                                  8
## 5          1 female      0-16          2012                                  6
## 6          1 female      0-16          2013                                  6
## Head of BITRE_172_children_0_16_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv :
##   SA4_CODE16 sex age_group year_range
## 1        101 all      0-16  2008-2017
## 2        101 all      0-16  2009-2018
## 3        101 all      0-16  2010-2019
## 4        101 all      0-16  2011-2020
## 5        101 all      0-16  2012-2021
## 6        101 all      0-16  2013-2022
##   rolling_average_fatally_injured_in_road_accident
## 1                                                3
## 2                                                4
## 3                                                4
## 4                                                4
## 5                                                5
## 6                                                5
## Head of BITRE_172_young_people_17_25_motor_vehicle_accidents_STE.csv :
##   STE_CODE16    sex age_group calendar_year type_of_road_user
## 1          1 female     17-25          2008            driver
## 2          1 female     17-25          2008             other
## 3          1 female     17-25          2008         passenger
## 4          1 female     17-25          2009            driver
## 5          1 female     17-25          2009             other
## 6          1 female     17-25          2009         passenger
##   n_fatally_injured_in_road_accident
## 1                                 13
## 2                                  1
## 3                                  6
## 4                                  7
## 5                                  3
```

```
## 6                                  10
## Head of BITRE_172_young_people_17_25_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv :
##   SA4_CODE16 sex age_group year_range type_of_road_user
## 1        101 all     17-25  2008-2017            driver
## 2        101 all     17-25  2008-2017         passenger
## 3        101 all     17-25  2008-2017             other
## 4        101 all     17-25  2009-2018            driver
## 5        101 all     17-25  2009-2018         passenger
## 6        101 all     17-25  2009-2018             other
##   rolling_average_fatally_injured_in_road_accident
## 1                                                3
## 2                                                0
## 3                                                3
## 4                                                4
## 5                                                1
## 6                                                3
```

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)
```

```
##
## BITRE_172_children_0_16_motor_vehicle_accidents_STE.csv
## BITRE_172_children_0_16_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv           BITRE_17
## BITRE_172_young_people_17_25_motor_vehicle_accidents_STE.csv
## BITRE_172_young_people_17_25_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv BITRE_172_you
##
## BITRE_172_children_0_16_motor_vehicle_accidents_STE.csv
## BITRE_172_children_0_16_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv
## BITRE_172_young_people_17_25_motor_vehicle_accidents_STE.csv                                    STE
## BITRE_172_young_people_17_25_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv SA4_CODE16, s
```

This code chunk iterates over each data frame in df_list, retrieves its variable names, class types, range for numeric variables, and unique values for character variables.

It then creates a data frame var_info with these details and prints it, providing a data dictionary for each data frame in the 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

```
# Iterate through each data frame and create a data dictionary
for (df_name in names(df_list)) {
  cat(paste("Data Dictionary for", df_name, ":\n"))

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

  # Create a data frame with variable name, class, range, unique values, and count of missing values
  var_info <- data.frame(
    variable = names(df),
    class = sapply(df, class),
```

```
    range = sapply(df, function(x) if (is.numeric(x)) paste(range(x, na.rm = TRUE), collapse = " - ") el
    unique_values = sapply(df, function(x) if (is.character(x)) paste(unique(x), collapse = ", ") else "
    n_missing_values = sapply(df, function(x) sum(is.na(x)))
  )

  # Print the data dictionary
  print(var_info)
}
```

```
## Data Dictionary for BITRE_172_children_0_16_motor_vehicle_accidents_STE.csv :
##                                                          variable     class
## STE_CODE16                                               STE_CODE16   integer
## sex                                                             sex character
## age_group                                                 age_group character
## calendar_year                                          calendar_year   integer
## n_fatally_injured_in_road_accident n_fatally_injured_in_road_accident   integer
##                                          range                 unique_values
## STE_CODE16                               1 - 8
## sex                                           female, male, NA, unspecified
## age_group                                                               0-16
## calendar_year                       2008 - 2022
## n_fatally_injured_in_road_accident      1 - 18
##                                    n_missing_values
## STE_CODE16                                        0
## sex                                               8
## age_group                                         0
## calendar_year                                     0
## n_fatally_injured_in_road_accident                0
## Data Dictionary for BITRE_172_children_0_16_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv
##                                                                                                  variable
## SA4_CODE16                                                                                     SA4_CODE16
## sex                                                                                                   sex
## age_group                                                                                      age_group
## year_range                                                                                     year_range
## rolling_average_fatally_injured_in_road_accident rolling_average_fatally_injured_in_road_accident
##                                                      class     range
## SA4_CODE16                                         integer 101 - 801
## sex                                              character
## age_group                                        character
## year_range                                       character
## rolling_average_fatally_injured_in_road_accident   integer    0 - 15
##                                                                                                    v
## SA4_CODE16
## sex
## age_group
## year_range                                       2008-2017, 2009-2018, 2010-2019, 2011-2020, 2012-202
## rolling_average_fatally_injured_in_road_accident
##                                                  n_missing_values
## SA4_CODE16                                                      0
## sex                                                             0
## age_group                                                       0
## year_range                                                      0
## rolling_average_fatally_injured_in_road_accident                0
```

```
## Data Dictionary for BITRE_172_young_people_17_25_motor_vehicle_accidents_STE.csv :
##                                                                variable     class
## STE_CODE16                                                   STE_CODE16   integer
## sex                                                                 sex character
## age_group                                                     age_group character
## calendar_year                                             calendar_year   integer
## type_of_road_user                                     type_of_road_user character
## n_fatally_injured_in_road_accident n_fatally_injured_in_road_accident   integer
##                                              range            unique_values
## STE_CODE16                                   1 - 8
## sex                                                          female, male
## age_group                                                           17-25
## calendar_year                          2008 - 2022
## type_of_road_user                                  driver, other, passenger
## n_fatally_injured_in_road_accident        1 - 40
##                                       n_missing_values
## STE_CODE16                                           0
## sex                                                  0
## age_group                                            0
## calendar_year                                        0
## type_of_road_user                                    0
## n_fatally_injured_in_road_accident                   0
## Data Dictionary for BITRE_172_young_people_17_25_rolling_sum_over_10_years_motor_vehicle_accidents_S
##                                                                                                  variable
## SA4_CODE16                                                                                     SA4_CODE16
## sex                                                                                                   sex
## age_group                                                                                        age_group
## year_range                                                                                      year_range
## type_of_road_user                                                                         type_of_road_user
## rolling_average_fatally_injured_in_road_accident rolling_average_fatally_injured_in_road_accident
##                                                        class     range
## SA4_CODE16                                           integer 101 - 801
## sex                                                character
## age_group                                          character
## year_range                                         character
## type_of_road_user                                  character
## rolling_average_fatally_injured_in_road_accident   integer    0 - 36
##                                                                                                    1
## SA4_CODE16
## sex
## age_group
## year_range                                         2008-2017, 2009-2018, 2010-2019, 2011-2020, 2012-201
## type_of_road_user                                                                        driver, pass
## rolling_average_fatally_injured_in_road_accident
##                                                    n_missing_values
## SA4_CODE16                                                        0
## sex                                                               0
## age_group                                                         0
## year_range                                                        0
## type_of_road_user                                                 0
## rolling_average_fatally_injured_in_road_accident                  0
```

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_val
      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
  }

  if (!is.na(first_col) && first_col != "Australia" && any(df[[1]] == 0, na.rm = TRUE)) {
  cat("Error: Values coded as 0 (Australia) found in a dataset that is not national:", df_name, "\n")
  }
}
```

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/bitre//"
output_dir <- "C:/Users/00095998/OneDrive - The University of Western Australia/acwa_temp/bitre//cell_su
```

Create the output directory if it doesn't already exist

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

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 BITRE_172_children_0_16_motor_vehi
## You don't have to worry about cell suppression on 'uncertainty' in BITRE_172_children_0_16_rolling_su
## You don't have to worry about cell suppression on 'uncertainty' in BITRE_172_young_people_17_25_motor
## You don't have to worry about cell suppression on 'uncertainty' in BITRE_172_young_people_17_25_rolli
```

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 BITRE_172_children_0_16_motor_vehicle_accidents_STE.csv, check values in column 'n_fatally_

## For file BITRE_172_children_0_16_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv, check val

## For file BITRE_172_young_people_17_25_motor_vehicle_accidents_STE.csv, check values in column 'n_fata

## For file BITRE_172_young_people_17_25_rolling_sum_over_10_years_motor_vehicle_accidents_SA4.csv, chec
```

Get a list of all CSV files in the input directory Loop through each CSV file and apply cell suppression

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

  # Check if the columns exist in the data frame
```

```r
  if ("n_fatally_injured_in_road_accident" %in% colnames(df) || "rolling_average_fatally_injured_in_roa
    suppressed_cols <- c("n_fatally_injured_in_road_accident", "rolling_average_fatally_injured_in_road_

    # Apply cell suppression and count affected rows
    affected_rows <- 0
    for (col in suppressed_cols) {
      if (col %in% colnames(df)) {
        suppressed_rows <- df[, col] < 5 & df[, col] > 0 & !is.na(df[, col])
        df[suppressed_rows, -c(1:4)] <- 9999999
        affected_rows <- affected_rows + sum(suppressed_rows)
      }
    }

    # Print the number of affected rows
    cat("Number of affected rows in", file, ":", affected_rows, "\n")

    # Print tables with variable information
    for (col in suppressed_cols) {
      if (col %in% colnames(df)) {
        cat("Variable:", col, "\n")
        table_df <- data.frame(
          Name = colnames(df),
          Range = sapply(df, function(x) paste(min(x, na.rm = TRUE), max(x, na.rm = TRUE), sep = "-")),
          Type = sapply(df, class),
          Unique_Values = sapply(df, function(x) length(unique(x))),
          Count_NA = sapply(df, function(x) sum(is.na(x)))
        )
        print(table_df)
        cat("\n")
      }
    }
  } else {
    # Print a message to indicate that the columns were not found
    cat("Skipping file", file, "because it does not contain the n_fatally_injured_in_road_accident or r
  }

  # 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)
}
```

```
## Number of affected rows in C:/Users/00095998/OneDrive - The University of Western Australia/acwa_temp
## Variable: n_fatally_injured_in_road_accident
##                                                                  Name
## STE_CODE16                                                 STE_CODE16
## sex                                                               sex
## age_group                                                   age_group
## calendar_year                                           calendar_year
## n_fatally_injured_in_road_accident n_fatally_injured_in_road_accident
##                                                 Range      Type Unique_Values
## STE_CODE16                                        1-8   integer             8
## sex                                 female-unspecified character             4
## age_group                                   0-16-0-16 character             1
## calendar_year                               2008-2022   integer            15
```

```
## n_fatally_injured_in_road_accident          5-9999999   numeric            13
##                                     Count_NA
## STE_CODE16                                 0
## sex                                        8
## age_group                                  0
## calendar_year                              0
## n_fatally_injured_in_road_accident         0
##
## Number of affected rows in C:/Users/00095998/OneDrive - The University of Western Australia/acwa_temp
## Variable: rolling_average_fatally_injured_in_road_accident
##                                                                                                  Name
## SA4_CODE16                                                                                  SA4_CODE16
## sex                                                                                            sex
## age_group                                                                                    age_group
## year_range                                                                                  year_range
## rolling_average_fatally_injured_in_road_accident rolling_average_fatally_injured_in_road_accident
##                                                                                    Range      Type
## SA4_CODE16                                                                       101-801   integer
## sex                                                                              all-all character
## age_group                                                                      0-16-0-16 character
## year_range                                                           2008-2017-2013-2022 character
## rolling_average_fatally_injured_in_road_accident                            0-9999999   numeric
##                                                                      Unique_Values Count_NA
## SA4_CODE16                                                                      78        0
## sex                                                                              1        0
## age_group                                                                        1        0
## year_range                                                                       6        0
## rolling_average_fatally_injured_in_road_accident                                13        0
##
## Number of affected rows in C:/Users/00095998/OneDrive - The University of Western Australia/acwa_temp
## Variable: n_fatally_injured_in_road_accident
##                                                                  Name
## STE_CODE16                                                 STE_CODE16
## sex                                                               sex
## age_group                                                   age_group
## calendar_year                                           calendar_year
## type_of_road_user                                     type_of_road_user
## n_fatally_injured_in_road_accident n_fatally_injured_in_road_accident
##                                                 Range      Type Unique_Values
## STE_CODE16                                        1-8   integer             8
## sex                                        female-male character             2
## age_group                                  17-25-17-25 character             1
## calendar_year                                2008-2022   integer            15
## type_of_road_user                      9999999-passenger character             4
## n_fatally_injured_in_road_accident           5-9999999   numeric            32
##                                     Count_NA
## STE_CODE16                                 0
## sex                                        0
## age_group                                  0
## calendar_year                              0
## type_of_road_user                          0
## n_fatally_injured_in_road_accident         0
##
## Number of affected rows in C:/Users/00095998/OneDrive - The University of Western Australia/acwa_temp
```

```
## Variable: rolling_average_fatally_injured_in_road_accident
##                                                                                               Name
## SA4_CODE16                                                                              SA4_CODE16
## sex                                                                                            sex
## age_group                                                                                age_group
## year_range                                                                              year_range
## type_of_road_user                                                                type_of_road_user
## rolling_average_fatally_injured_in_road_accident rolling_average_fatally_injured_in_road_accident
##                                                               Range      Type
## SA4_CODE16                                                  101-801   integer
## sex                                                         all-all character
## age_group                                               17-25-17-25 character
## year_range                                      2008-2017-2013-2022 character
## type_of_road_user                                 9999999-passenger character
## rolling_average_fatally_injured_in_road_accident          0-9999999   numeric
##                                                  Unique_Values Count_NA
## SA4_CODE16                                                  88        0
## sex                                                          1        0
## age_group                                                    1        0
## year_range                                                   6        0
## type_of_road_user                                            4        0
## rolling_average_fatally_injured_in_road_accident            29        0
```