

MATH2349 Semester 1, 2018

[Code ▾](#)

Assignment 1 - Victorian family violence cases 2012-2017

Phil Steinke s3725547@student.rmit.edu.au (<mailto:s3725547@student.rmit.edu.au>)

Setup

Data Description

"Victims Support Agency Data Tables- 2016-17.xlsx" Table 2. Number of VAP family violence cases initiated for new clients by client gender and age group, July 2012 to June 2017

Source:

https://www.crimestatistics.vic.gov.au/sites/default/files/embridge_cache/emshare/original/public/2017/12/74/906ab3fb8/Victims%20Support%20Agency%20Data%202016-17.xlsx

(https://www.crimestatistics.vic.gov.au/sites/default/files/embridge_cache/emshare/original/public/2017/12/74/906ab3fb8/Victims%20Support%20Agency%20Data%202016-17.xlsx)

[Hide](#)

```
#' VICTIMS SUPPORT AGENCY DATA TABLES- 2016-17.XLSX
#' TABLE 2. NUMBER OF VAP FAMILY VIOLENCE CASES INITIATED FOR NEW CLIENTS BY CLIENT GENDER AND AGE GROUP,
#' JULY 2012 TO JUNE 2017
#'
#' VAP FAMILY VIOLENCE CASES INITIATED FOR NEW CLIENTS BY CLIENT GENDER AND AGE GROUP
#'
#' @format Starting format xlsx with 53 observations/rows and 7 variables/cols
#' only 46 observations/rows and 7 variables/cols are imported to not include whitespace and the totals
#' \describe{
#'   \item{\code{Gender}}{character. GENDER OF CLIENT THAT REPORTED FAMILY VIOLENCE. LEVELS: MALE/FEMALE.}
#'   \item{\code{age group}}{character. DESCRIPTION.}
#'   \item{\code{2012-13}}{character. COUNT OF REPORTED INCIDENTS FROM JULY 2012 TILL JUNE 2013.}
#'   \item{\code{2013-14}}{character. COUNT OF REPORTED INCIDENTS FROM JULY 2013 TILL JUNE 2014.}
#'   \item{\code{2014-15}}{character. COUNT OF REPORTED INCIDENTS FROM JULY 2014 TILL JUNE 2015.}
#'   \item{\code{2015-16}}{character. COUNT OF REPORTED INCIDENTS FROM JULY 2015 TILL JUNE 2016.}
#'   \item{\code{2016-17}}{character. COUNT OF REPORTED INCIDENTS FROM JULY 2016 TILL JUNE 2017.}
#' }
#' "Victims Support Agency Data Tables- 2016-17.xlsx"
#'
#' @format end format is a dataframe with 28 observations/rows and 7 variables/cols
#' \describe{
#'   \item{\code{Gender}}{character. GENDER OF CLIENT THAT REPORTED FAMILY VIOLENCE. LEVELS: MALE/FEMALE. }
#'   \item{\code{Age Range}}{character. AGE OF PARTICIPANTS DIVIDED INTO 5 YEAR INCRIMENTS.}
#'   \item{\code{2012-13}}{integer. COUNT OF REPORTED INCIDENTS FROM JULY 2012 TILL JUNE 2013.}
#'   \item{\code{2013-14}}{integer. COUNT OF REPORTED INCIDENTS FROM JULY 2013 TILL JUNE 2014.}
#'   \item{\code{2014-15}}{integer. COUNT OF REPORTED INCIDENTS FROM JULY 2014 TILL JUNE 2015.}
#'   \item{\code{2015-16}}{integer. COUNT OF REPORTED INCIDENTS FROM JULY 2015 TILL JUNE 2016.}
#'   \item{\code{2016-17}}{integer. COUNT OF REPORTED INCIDENTS FROM JULY 2016 TILL JUNE 2017.}
#' }
"family_violence"
```

```
[1] "family_violence"
```

As a minimum, your data set should include: * one numeric variable = number of family violence cases per year * one qualitative (categorical) variable = Age Range

This dataset show most reported assaults with women occur between the ages of 25-49. Each age bracket within that range (of 5 years) have approximately double the reported assaults of children and teenagers. The data also shows an increase in reported assaults per year over the last 5 years.

Read/Import Data

[Hide](#)

```
rm(list=ls())
setwd("~/code/tldr/data-science/data-preprocessing-math2349/assignment1/data/")
```

The working directory was changed to /Users/phil/code/tldr/data-science/data-preprocessing-math2349/assignment1/data inside a notebook chunk. The working directory will be reset when the chunk is finished running. Use the `knitr::setwd()` option in the setup chunk to change the working directory for notebook chunks.

[Hide](#)

```
# Read/Import the data into R, then save it as a data frame.
family_violence <-
  read_excel(
    "Victims Support Agency Data Tables- 2016-17.xlsx",
    sheet = "Table 2",
    range = cell_rows(12:58)
  ) %>%
  data.frame()
# `stringsAsFactors = FALSE` wont work here, so I set it in my defaults
class(family_violence) # -> family violence is a "data.frame"
```

```
[1] "data.frame"
```

Hide

```
# You must also provide the R codes with outputs
head(family_violence)
```

X_1 <chr>	X_2 <chr>	X_3 <chr>	X_4 <chr>	X_5 <chr>	X_6 <chr>	X_7 <chr>
1 Gender and age group	NA	2012-13	2013-14	2014-15	2015-16	2016-17
2 Male	0 - 4	74	61	63	41	29
3 NA	5 - 9	84	121	120	107	97
4 NA	10 - 14	72	80	107	88	100
5 NA	15 - 19	52	70	82	74	80
6 NA	20 - 24	64	47	74	95	90

6 rows

Inspect and Understand

* check the dimensions of the data frame.

Hide

```
dim(family_violence)
```

```
[1] 46  7
```

Hide

```
# OR
nrow(family_violence)
```

```
[1] 46
```

Hide

```
ncol(family_violence)
```

```
[1] 7
```

Hide

```
# check the attributes in the data.
attributes(family_violence)
```

```
$names
[1] "X__1" "X__2" "X__3" "X__4" "X__5" "X__6" "X__7"

$row.names
[1] 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
[36] 36 37 38 39 40 41 42 43 44 45 46

$class
[1] "data.frame"
```

- It has 46 rows and 7 columns
- It's names are X__1, ...
- It's row names are numbers 1,2,3...

- It's a data.frame

check the data types (i.e., character, numeric, integer, factor, and logical) of the variables in the data set.

Hide

```
class(family_violence) # -> family_violence is a data.frame
```

```
[1] "data.frame"
```

Hide

```
class(family_violence[,1]) # -> "Gender and age group" is a character
```

```
[1] "character"
```

Hide

```
head(family_violence[,2]) # [1] NA          "0 - 4"      "5 - 9"      "10 - 14"    "15 - 19"    "20 - 24"
```

```
[1] NA          "0 - 4"      "5 - 9"      "10 - 14"    "15 - 19"    "20 - 24"
```

Hide

```
class(family_violence[,2]) # -> is a character
```

```
[1] "character"
```

Hide

```
head(family_violence[,3]) # [1] "2012-13" "74"      "84"      "72"      "52"      "64"      "62"
```

```
[1] "2012-13" "74"      "84"      "72"      "52"      "64"
```

Hide

```
class(family_violence[,3])# -> is a character
```

```
[1] "character"
```

Hide

```
head(family_violence[,4]) # [1] "2013-14" "61"      "121"     "80"      "70"      "47"      "60"
```

```
[1] "2013-14" "61"      "121"     "80"      "70"      "47"
```

Hide

```
class(family_violence[,4]) # -> is a character
```

```
[1] "character"
```

Hide

```
head(family_violence[,5]) # [1] "2013-14" "61"      "121"     "80"      "70"      "47"      "60"
```

```
[1] "2014-15" "63"      "120"     "107"     "82"      "74"
```

Hide

```
class(family_violence[,5]) # -> is a character
```

```
[1] "character"
```

Hide

```
head(family_violence[,6]) # "2015-16" "41"      "107"     "88"      "74"      "95"
```

```
[1] "2015-16" "41" "107" "88" "74" "95"
```

Hide

```
class(family_violence[,6]) # -> is a character
```

```
[1] "character"
```

- Everything is treated as a character because of the column titles are included in the spreadsheet

check the levels of factor variables

Hide

```
# family_violence[,1] # column names for reference
levels_gender <-
  c(family_violence[,1]) %>%
  factor(ordered= TRUE) %>%
  levels() %>%
  print()
```

```
[1] "Female" "Gender and age group" "Male" "Total persons2"
```

Hide

```
levels_age_range <-
  c(family_violence[,2]) %>%
  factor(ordered= TRUE) %>%
  levels() %>%
  print()
```

```
[1] "0 - 4" "10 - 14" "15 - 19" "20 - 24" "25 - 29" "30 - 34" "35 - 39"
[8] "40 - 44" "45 - 49" "5 - 9" "50 - 54" "55 - 59" "60 - 64" "65 and older"
[15] "Total1"
```

Hide

```
cat("\nLevels for all year cols from 2012-17\n including titles")
```

```
Levels for all year cols from 2012-17
including titles
```

Hide

```
levels_all_years <-
c(family_violence[,3],
  family_violence[,4],
  family_violence[,5],
  family_violence[,6],
  family_violence[,7]
) %>%
  factor() %>%
  levels() %>%
  print()
```

```
[1] "100" "101" "103" "104" "105" "106" "107" "110" "112" "113"
[11] "114" "117" "118" "120" "121" "122" "1227" "1234" "126" "128"
[21] "129" "134" "136" "139" "140" "141" "142" "1451" "148" "149"
[31] "153" "1532" "154" "162" "163" "166" "167" "173" "188" "189"
[41] "190" "191" "195" "1955" "201" "2012-13" "2013-14" "2014-15" "2015-16" "2016-17"
[51] "202" "203" "207" "210" "214" "215" "217" "2181" "219" "221"
[61] "222" "223" "224" "225" "226" "229" "230" "231" "232" "238"
[71] "240" "241" "2433" "244" "2444" "247" "254" "260" "262" "268"
[81] "269" "2735" "278" "28" "280" "286" "288" "289" "29" "290"
[91] "2909" "294" "296" "305" "310" "318" "321" "334" "336" "338"
[101] "357" "360" "361" "368" "3680" "3727" "376" "38" "388" "393"
[111] "397" "3987" "400" "41" "42" "423" "43" "432" "433" "44"
[121] "45" "47" "48" "51" "52" "53" "54" "57" "59" "60"
[131] "61" "62" "63" "64" "65" "66" "67" "68" "69" "70"
[141] "71" "72" "74" "75" "76" "77" "78" "80" "82" "84"
[151] "85" "86" "87" "88" "89" "90" "91" "93" "94" "947"
[161] "95" "96" "97" "984" "99"
```

Hide

```
cat("\nLevels from 2012-13\n")
```

```
Levels from 2012-13
```

[Hide](#)

```
levels_2012_13 <-  
  c(family_violence[,3]) %>%  
  factor(ordered= TRUE) %>%  
  levels() %>%  
  print()
```

```
[1] "103" "122" "128" "139" "1451" "149" "154" "162" "163" "188"  
[11] "191" "201" "2012-13" "232" "2444" "269" "280" "29" "38" "42"  
[21] "44" "47" "52" "59" "62" "64" "65" "66" "68" "70"  
[31] "71" "72" "74" "76" "77" "84" "87" "89" "984"
```

* check the column names in the data frame, rename them if required.

[Hide](#)

```
# check the column names in the data frame  
colnames(family_violence)
```

```
[1] "X_1" "X_2" "X_3" "X_4" "X_5" "X_6" "X_7"
```

[Hide](#)

```
# rename them if required.  
colnames(family_violence) <- c("Gender", "Age Range", c(family_violence[1,3:7]))  
#The excel doesn't include Male/Female accross all of the fields, so here I've filled them in:  
family_violence[c(3:16),1] <- "Male"  
family_violence[c(18:31),1] <- "Female"  
# Removing the empty rows and rows with totals in them  
family_violence <- family_violence[-c(1, 16, 31:46), ]  
# Fixing the Row numbering  
rownames(family_violence) <- c(1:length(family_violence$`Gender`))  
family_violence
```

	Gender <chr>	Age Range <chr>	2012-13 <chr>	2013-14 <chr>	2014-15 <chr>	2015-16 <chr>	2016-17 <chr>
1	Male	0 - 4	74	61	63	41	29
2	Male	5 - 9	84	121	120	107	97
3	Male	10 - 14	72	80	107	88	100
4	Male	15 - 19	52	70	82	74	80
5	Male	20 - 24	64	47	74	95	90
6	Male	25 - 29	62	60	100	101	112
7	Male	30 - 34	70	61	78	91	120
8	Male	35 - 39	68	67	80	82	148
9	Male	40 - 44	87	78	91	114	134
10	Male	45 - 49	65	72	105	114	163
1-10 of 28 rows							Previous 1 2 3 Next

[Hide](#)

```
class(family_violence) # -> family_violence is a data.frame
```

```
[1] "data.frame"
```

[Hide](#)

```
family_violence[1, 'Age Range'] # -> "0 - 4"
```

```
[1] "0 - 4"
```

Hide

```
class(family_violence[3, 'Age Range']) # -> "Age Range" is a character
```

```
[1] "character"
```

Hide

```
family_violence[3, '2012-13']
```

```
[1] "72"
```

Hide

```
class(family_violence[3, '2012-13']) # -> "Year" is a character
```

```
[1] "character"
```

Hide

```
family_violence[1, 4]
```

```
[1] "61"
```

Hide

```
class(family_violence[1, 4]) # -> "Gender" and N/A is a character
```

```
[1] "character"
```

Hide

```
# fixing the data types: rename/rearrange if required  
cat("Setting each year's data to integers\n")
```

```
Setting each year's data to integers
```

Hide

```
class(family_violence[3:7])
```

```
[1] "data.frame"
```

Hide

```
family_violence[3:7] <- Map(as.integer, family_violence[3:7])  
Map(is.integer, family_violence[3:7])
```

```
$`2012-13`  
[1] TRUE
```

```
$`2013-14`  
[1] TRUE
```

```
$`2014-15`  
[1] TRUE
```

```
$`2015-16`  
[1] TRUE
```

```
$`2016-17`  
[1] TRUE
```

Hide

```
# Previous code that seemed cumbersome:  
#class(family_violence$`2012-13`)  
family_violence$`2012-13` %>%  
  as.integer() -> family_violence$`2012-13`  
#class(family_violence$`2012-13`)  
cat("\nLevels for all years again: from 2012-17\n including titles")
```

Levels for all years again: from 2012-17
including titles

Hide

```
levels_all_years <-  
c(family_violence[,3],  
  family_violence[,4],  
  family_violence[,5],  
  family_violence[,6],  
  family_violence[,7]  
) %>%  
  factor() %>%  
  levels() %>%  
  print()
```

```
[1] "28" "29" "38" "41" "42" "43" "44" "45" "47" "48" "51" "52" "53" "54" "57" "59" "60"  
[18] "61" "62" "63" "64" "65" "66" "67" "68" "69" "70" "72" "74" "75" "77" "78" "80" "82"  
[35] "84" "85" "86" "87" "88" "89" "90" "91" "93" "94" "95" "96" "97" "99" "100" "101" "105"  
[52] "106" "107" "110" "112" "113" "114" "118" "120" "121" "122" "126" "128" "129" "134" "136" "141" "142"  
[69] "148" "154" "162" "163" "166" "167" "189" "191" "195" "201" "202" "210" "219" "222" "225" "226" "229"  
[86] "231" "238" "244" "247" "260" "262" "268" "288" "294" "310" "318" "321" "336"
```

New data types tests

Hide

```
cat("New data types\n")
```

New data types

Hide

```
class(family_violence) # -> family_violence is a data.frame
```

```
[1] "data.frame"
```

Hide

```
cat("Age Range\n")
```

Age Range

Hide

```
family_violence[1, 'Age Range'] # -> "0 - 4"
```

```
[1] "0 - 4"
```

Hide

```
class(family_violence[3, 'Age Range']) # -> "Age Range" is a character
```

```
[1] "character"
```

Hide

```
cat("Year col 2012-13\n")
```

Year col 2012-13

Hide

```
family_violence$'2012-13'
```

```
[1] 74 84 72 52 64 62 70 68 87 65 77 38 42 59 47 65 66 70 89 128 162 201 191 122 84 38  
[27] 29 44
```

Hide

```
class(family_violence$'2012-13') # -> All Year cols are now an integer
```

```
[1] "integer"
```

Hide

```
cat("single value from a year column 2012-13\n")
```

```
single value from a year column 2012-13
```

Hide

```
family_violence[1, 5]
```

```
[1] 63
```

Hide

```
class(family_violence[1, 4]) # -> Grabbing a single value from a year col which is now an integer
```

```
[1] "integer"
```

Hide

```
dim(family_violence)
```

```
[1] 28  7
```

Subsetting I

Subset the data frame using first 10 observations (include all variables). Then convert it to a matrix.

Hide

```
# Subset the data frame using first 10 observations (include all variables)
# What are all variables?
names(family_violence) -> all_variables
all_variables
```

```
[1] "Gender"      "Age Range" "2012-13"   "2013-14"   "2014-15"   "2015-16"   "2016-17"
```

Hide

```
# I assume you mean this because all_variables
data_frame_subset <- family_violence[1:10,]
data_frame_subset
```

	Gender <chr>	Age Range <chr>	2012-13 <int>	2013-14 <int>	2014-15 <int>	2015-16 <int>	2016-17 <int>
1	Male	0 - 4	74	61	63	41	29
2	Male	5 - 9	84	121	120	107	97
3	Male	10 - 14	72	80	107	88	100
4	Male	15 - 19	52	70	82	74	80
5	Male	20 - 24	64	47	74	95	90
6	Male	25 - 29	62	60	100	101	112
7	Male	30 - 34	70	61	78	91	120
8	Male	35 - 39	68	67	80	82	148
9	Male	40 - 44	87	78	91	114	134
10	Male	45 - 49	65	72	105	114	163

1-10 of 10 rows

Hide

```
# Then convert it to a matrix
data_frame_subset %>%
  as.matrix(
  ) %>%
  print()
```



```

      Gender Age Range 2012-13 2013-14 2014-15 2015-16 2016-17
1  "Male"  "0 - 4"   "74"    " 61"    " 63"    " 41"    " 29"
2  "Male"  "5 - 9"   "84"    "121"    "120"    "107"    " 97"
3  "Male" "10 - 14"  "72"    " 80"    "107"    " 88"    "100"
4  "Male" "15 - 19"  "52"    " 70"    " 82"    " 74"    " 80"
5  "Male" "20 - 24"  "64"    " 47"    " 74"    " 95"    " 90"
6  "Male" "25 - 29"  "62"    " 60"    "100"    "101"    "112"
7  "Male" "30 - 34"  "70"    " 61"    " 78"    " 91"    "120"
8  "Male" "35 - 39"  "68"    " 67"    " 80"    " 82"    "148"
9  "Male" "40 - 44"  "87"    " 78"    " 91"    "114"    "134"
10 "Male" "45 - 49"  "65"    " 72"    "105"    "114"    "163"

```

Hide

```
data_frame_matrix1 <- data.matrix(data_frame_subset, rownames.force = NA)
```

```
NAS introduced by coercionNAS introduced by coercion
```

Hide

```
class(data_frame_matrix1) # Matrix
```

```
[1] "matrix"
```

Hide

```
data_frame_matrix2 <- as.matrix(data_frame_subset)
class(data_frame_matrix2) # Matrix
```

```
[1] "matrix"
```

Hide

```
data_frame_matrix3 <- apply(data_frame_subset, 2, as.matrix)
class(data_frame_matrix3) # Matrix Trinity
```

```
[1] "matrix"
```

Subsetting II

Hide

```
## Subset the data frame including only first and the last variable in the data set
# Grabbing the variables:
names(family_violence) -> all_variables
all_variables
```

```
[1] "Gender"      "Age Range"  "2012-13"    "2013-14"    "2014-15"    "2015-16"    "2016-17"
```

Hide

```
family_violence %>%
  subset (
    select = c(
      1,
      length(family_violence)
    )
  ) -> first_and_last_subset
head(first_and_last_subset)
```

	Gender <chr>	2016-17 <int>
1	Male	29
2	Male	97
3	Male	100
4	Male	80
5	Male	90
6	Male	112
6 rows		

save it as an R object file (.RData).

This didn't work:

Hide

```
save.image() # Saving the workspace
first_and_last_subset
```

	Gender <chr>	2016-17 <int>
1	Male	29
2	Male	97
3	Male	100
4	Male	80
5	Male	90
6	Male	112
7	Male	120
8	Male	148
9	Male	134
10	Male	163
1-10 of 28 rows		Previous 1 2 3 Next

Hide

```
save(first_and_last_subset, file = "data/first_and_last_subset.Rdata")
rm(first_and_last_subset)
testing_save_worked <- load("data/first_and_last_subset.Rdata")
identical(first_and_last_subset, testing_save_worked) # FALSE
```

```
[1] FALSE
```

Hide

```
# Using load.Rdata2 from miceadds instead:
save.Rdata(first_and_last_subset, "data/first_and_last_subset.RData")
testing_save_worked <- load.Rdata2(filename = "data/first_and_last_subset.RData", path=getwd())
identical(first_and_last_subset, testing_save_worked) # [1] TRUE
```

```
[1] TRUE
```