Tables

# Statement of purpose

Various R packages aiming to facilitate the creation of descriptive and modelling summary tables have seen light in recent years. Tabulation has for a long time been a limitation in base R, particularly in respect to creating the types of tables that summarise typical social scientific variables and datasets (i.e. categorical variables with labelled values). Various recent packages now facilitate the creation of tables, focusing either on (a) creating near-to-publishable quality summary tables for specific purposes with very little and easy code, or (b) allowing maximum flexibility to create any table imaginable with more complex code. However, it is harder to find functions that achieve a good middle ground between the two. Furthermore, most of the table functions work best for HTML output or - with the additional complexities introduced by LaTex - PDF. Options for output to Microsoft Word (and related popular text editing) format is severely limited.

My aim here is to identify the optimal packages and functions for all tabulation needs that satisfy the following criteria, primarily with the purpose of facilitating teaching:

* the fewest packages needed to rely upon;
* the simplest code;
* code that follows the same pattern as other functions used for other purposes (e.g. formula syntax as with modelling; graphics syntax, as with ggplot2);
* prints well to MS Word (or produces an object type that can easily be passed on to {flextable} or similar)

## Packages

library(here)  
library(tidyverse)  
library(janitor)  
library(easystats) # datawizard  
library(modelsummary)  
library(strengejacke)

# Attaching packages  
✔ ggeffects 1.3.1 ✔ sjlabelled 1.2.0   
✔ sjmisc 2.8.9 ✔ sjstats 0.18.2  
✔ sjPlot 2.8.15 ✔ esc 0.5.1

library(gt)  
library(flextable)

## Data

# data1 <- readRDS(here("Data", "for\_analysis", "wvs7.rds")) |>  
data <- data\_read(here("Data", "for\_analysis", "wvs7.rds")) |>  
 sjlabelled::drop\_labels() |>  
 data\_extract(c(B\_COUNTRY, Q1, Q57, Q195, Q260, Q261:Q263)) |>  
 mutate(B\_COUNTRY = as\_character(B\_COUNTRY),  
 across(Q261:Q262, as\_numeric)) |>   
 sample\_n(1000)

## Dataset summaries

## gt\_preview  
  
gt::gt\_preview(data)

|  | B\_COUNTRY | Q1 | Q57 | Q195 | Q260 | Q261 | Q262 | Q263 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Iran | Very important | Need to be very careful | Never justifiable | Female | 1977 | 41 | I am born in this country |
| 2 | Germany | Very important | Need to be very careful | 8 | Female | 1992 | 26 | I am born in this country |
| 3 | Andorra | Very important | Need to be very careful | 6 | Female | 1989 | 30 | I am born in this country |
| 4 | Brazil | Very important | Most people can be trusted | Never justifiable | Male | 1970 | 49 | I am an immigrant to this country (born outside this country) |
| 5 | Puerto Rico | Very important | Need to be very careful | Never justifiable | Female | 1993 | 24 | I am born in this country |
| 6..999 |  |  |  |  |  |  |  |  |
| 1000 | Nicaragua | Very important | Need to be very careful | 9 | Male | 1988 | 30 | I am born in this country |

data |>   
 data\_codebook() |> print\_html(variable\_label\_width = 20,  
 value\_label\_width = 20,  
 font\_size = "80%", line\_padding = 0)

Table 1: data (1000 rows and 8 variables, 8 shown)

| ID | Name | Label | Type | Missings | Values | N |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | B\_COUNTRY | ISO 3166-1 numeric country code | character | 0 (0.0%) | Andorra | 12 (1.2%) |
|  |  |  |  |  | Argentina | 40 (4.0%) |
|  |  |  |  |  | Armenia | 10 (1.0%) |
|  |  |  |  |  | Australia | 11 (1.1%) |
|  |  |  |  |  | Bangladesh | 32 (3.2%) |
|  |  |  |  |  | Bolivia | 10 (1.0%) |
|  |  |  |  |  | Brazil | 14 (1.4%) |
|  |  |  |  |  | Canada | 10 (1.0%) |
|  |  |  |  |  | Chile | 12 (1.2%) |
|  |  |  |  |  | China | 9 (0.9%) |
|  |  |  |  |  | (...) |  |
| 2 | Q1 | Important in life: Family | categorical | 1 (0.1%) | Very important | 901 (90.2%) |
|  |  |  |  |  | Rather important | 87 (8.7%) |
|  |  |  |  |  | Not very important | 10 (1.0%) |
|  |  |  |  |  | Not at all important | 1 (0.1%) |
| 3 | Q57 | Most people can be trusted | categorical | 13 (1.3%) | Most people can be trusted | 250 (25.3%) |
|  |  |  |  |  | Need to be very careful | 737 (74.7%) |
| 4 | Q195 | Justifiable: Death penalty | categorical | 24 (2.4%) | Never justifiable | 361 (37.0%) |
|  |  |  |  |  | 2 | 72 (7.4%) |
|  |  |  |  |  | 3 | 68 (7.0%) |
|  |  |  |  |  | 4 | 34 (3.5%) |
|  |  |  |  |  | 5 | 121 (12.4%) |
|  |  |  |  |  | 6 | 76 (7.8%) |
|  |  |  |  |  | 7 | 41 (4.2%) |
|  |  |  |  |  | 8 | 63 (6.5%) |
|  |  |  |  |  | 9 | 30 (3.1%) |
|  |  |  |  |  | Always justifiable | 110 (11.3%) |
| 5 | Q260 | Sex | categorical | 2 (0.2%) | Male | 487 (48.8%) |
|  |  |  |  |  | Female | 511 (51.2%) |
| 6 | Q261 | Year of birth | numeric | 18 (1.8%) | [1921, 2003] | 982 |
| 7 | Q262 | Age | numeric | 4 (0.4%) | [17, 99] | 996 |
| 8 | Q263 | Respondent immigrant | categorical | 9 (0.9%) | I am born in this country | 930 (93.8%) |
|  |  |  |  |  | I am an immigrant to this country (born outside this country) | 61 (6.2%) |
|  | | | | | | |

## One-way tabulations (frequencies tables)

#### Base R

# Base R

#### {tidyverse} dplyr::count()

# {tidyverse} dplyr::count()  
  
data |> count(Q263) |> # basic: N's only  
 mutate(`%` = (n / sum(n)) \* 100) # compute % by hand

Q263 n %  
1 I am born in this country 930 93.0  
2 I am an immigrant to this country (born outside this country) 61 6.1  
3 <NA> 9 0.9

# janitor::adorn\_percentages("col")  
 # janitor::adorn\_totals("row") # add Totals using `janitor`

#### janitor::tabyl()

data |> tabyl(Q263) |>   
 flextable()

| Q263 | n | percent | valid\_percent |
| --- | --- | --- | --- |
| I am born in this country | 930 | 0.930 | 0.93844601 |
| I am an immigrant to this country (born outside this country) | 61 | 0.061 | 0.06155399 |
|  | 9 | 0.009 |  |

#### {easystats} datawizard::data\_tabulate()

data |> datawizard::data\_tabulate(Q263) |> sjmisc::print\_html() # needs help from sjmisc to print in Word

Table 1: Respondent immigrant (Q263) (categorical)

| Value | N | Raw % | Valid % | Cumulative % |
| --- | --- | --- | --- | --- |
| I am born in this country | 930 | 93.00 | 93.84 | 93.84 |
| I am an immigrant to this country (born outside this country) | 61 | 6.10 | 6.16 | 100.00 |
| (NA) | 9 | 0.90 | (NA) | (NA) |
| total N=1000 valid N=991 | | | | |

data |> datawizard::data\_tabulate(Q263) |> sjmisc::print\_md()

Respondent immigrant (Q263) (categorical)

| Value | N | Raw % | Valid % | Cumulative % |
| --- | --- | --- | --- | --- |
| I am born in this country | 930 | 93.00 | 93.84 | 93.84 |
| I am an immigrant to this country (born outside this country) | 61 | 6.10 | 6.16 | 100.00 |
| (NA) | 9 | 0.90 | (NA) | (NA) |

total N=1000 valid N=991

#### modelsummary::datasummary()

# Needs formula syntax  
  
  
modelsummary::datasummary(Q57 ~ N, data = data) # most basic: N only; prints to Word

| Q57 | N |
| --- | --- |
| Most people can be trusted | 250 |
| Need to be very careful | 737 |

modelsummary::datasummary(Q57 + 1 ~ N + Percent() + 1, data = data) # most complex: N, % and full marginals

| Q57 | N | Percent | All |
| --- | --- | --- | --- |
| Most people can be trusted | 250 | 25.00 | 250.00 |
| Need to be very careful | 737 | 73.70 | 737.00 |
| All | 1000 | 100.00 | 1000.00 |

modelsummary::datasummary(Q57 + 1 ~ N + Percent(), # Column marginals only  
 data = data,   
 output = "markdown") # Nicer print to Word within rmd/qmd

| Q57 | N | Percent |
| --- | --- | --- |
| Most people can be trusted | 250 | 25.00 |
| Need to be very careful | 737 | 73.70 |
| All | 1000 | 100.00 |

#### gtsummary::tbl\_summary()

data |> gtsummary::tbl\_summary(include = Q57) # a bit hacky; meant to summarise dataframes, not single variables

| **Characteristic** | **N = 1,000**1 |
| --- | --- |
| Most people can be trusted |  |
| Most people can be trusted | 250 (25%) |
| Need to be very careful | 737 (75%) |
| Unknown | 13 |
| 1n (%) | |

# , missing = "no"  
# |> gtsummary::modify\_header(label ~ "")

## Two-way tabulations (corsstabulations)

#### Base R

# Base R

#### {tidyverse} dplyr::

# {tidyverse} dplyr::  
  
data |> count(Q57, Q263)

Q57  
1 Most people can be trusted  
2 Most people can be trusted  
3 Most people can be trusted  
4 Need to be very careful  
5 Need to be very careful  
6 Need to be very careful  
7 <NA>  
8 <NA>  
9 <NA>  
 Q263 n  
1 I am born in this country 224  
2 I am an immigrant to this country (born outside this country) 23  
3 <NA> 3  
4 I am born in this country 697  
5 I am an immigrant to this country (born outside this country) 36  
6 <NA> 4  
7 I am born in this country 9  
8 I am an immigrant to this country (born outside this country) 2  
9 <NA> 2

#### {easystats} datawizard::data\_tabulate()

# `datawizard`  
  
# by default creates `n\_distinct(data$B\_COUNTRY)` number of separate frequency tables of `data$Q263`  
# the `collapse` option collapses the output into one table  
  
data |>   
 # group\_by(B\_COUNTRY) |>  
 data\_group(Q260) |>  
 data\_tabulate(Q263, collapse = TRUE) # |> print\_md()

#### janitor::tabyl()

data |> tabyl(Q57, Q263, show\_na = FALSE) |>   
 adorn\_totals(c("row", "col")) |>  
 adorn\_percentages("row") |>   
 adorn\_pct\_formatting(rounding = "half up", digits = 1) |>   
 adorn\_ns() |>   
 flextable()

| Q57 | I am born in this country | I am an immigrant to this country (born outside this country) | Total |
| --- | --- | --- | --- |
| Most people can be trusted | 90.7% (224) | 9.3% (23) | 100.0% (247) |
| Need to be very careful | 95.1% (697) | 4.9% (36) | 100.0% (733) |
| Total | 94.0% (921) | 6.0% (59) | 100.0% (980) |

### gtsummary::tbl\_cross()

data |>   
 gtsummary::tbl\_cross(Q57, Q260,  
 percent = "row",  
 missing = "no") # "no", "always", "ifany"

|  | Sex | |  |
| --- | --- | --- | --- |
|  | Male | Female | Total |
| Most people can be trusted |  |  |  |
| Most people can be trusted | 121 (48%) | 129 (52%) | 250 (100%) |
| Need to be very careful | 359 (49%) | 377 (51%) | 736 (100%) |
| Total | 480 (49%) | 506 (51%) | 986 (100%) |

### modelsummary::datasummary\_crosstab()

Major shortcoming that the “All” total marginals summarise the entire data, with the missing values of the crosstabulated variables included, while the missign values themselves are not shown within the table, so it can be very confusing and dangerous.

modelsummary::datasummary(Q57 + 1 ~ Q260 + Percent("col") + 1, data = data) # limited options for adding column and row statistics

| Q57 | Male | Female | Percent | All |
| --- | --- | --- | --- | --- |
| Most people can be trusted | 121.00 | 129.00 | 25.00 | 250.00 |
| Need to be very careful | 359.00 | 377.00 | 73.70 | 737.00 |
| All | 487.00 | 511.00 | 100.00 | 1000.00 |

modelsummary::datasummary\_crosstab(Q57 ~ Q260, data = data) # relatively good defaults; BUT "All" includes missing values without explicitly telling

| Q57 |  | Male | Female | All |
| --- | --- | --- | --- | --- |
| Most people can be trusted | N | 121 | 129 | 250 |
|  | % row | 48.4 | 51.6 | 100.0 |
| Need to be very careful | N | 359 | 377 | 737 |
|  | % row | 48.7 | 51.2 | 100.0 |
| All | N | 487 | 511 | 1000 |
|  | % row | 48.7 | 51.1 | 100.0 |

modelsummary::datasummary\_crosstab(Q57 ~ Q260, data = na.omit(data[c("Q57", "Q260")])) # This is what we would expect

| Q57 |  | Male | Female | All |
| --- | --- | --- | --- | --- |
| Most people can be trusted | N | 121 | 129 | 250 |
|  | % row | 48.4 | 51.6 | 100.0 |
| Need to be very careful | N | 359 | 377 | 736 |
|  | % row | 48.8 | 51.2 | 100.0 |
| All | N | 480 | 506 | 986 |
|  | % row | 48.7 | 51.3 | 100.0 |

modelsummary::datasummary\_crosstab(Q57 ~ Q260, statistic = 1 ~ N + Percent("col") + Percent("row") + 1, data = data) # can add/remove statistics

| Q57 |  | Male | Female | All |
| --- | --- | --- | --- | --- |
| Most people can be trusted | N | 121 | 129 | 250 |
|  | % col | 24.8 | 25.2 | 25.0 |
|  | % row | 48.4 | 51.6 | 100.0 |
| Need to be very careful | N | 359 | 377 | 737 |
|  | % col | 73.7 | 73.8 | 73.7 |
|  | % row | 48.7 | 51.2 | 100.0 |
| All | N | 487 | 511 | 1000 |
|  | % col | 100.0 | 100.0 | 100.0 |
|  | % row | 48.7 | 51.1 | 100.0 |

# Doesn't render to Word; doesn't work well with piping  
  
with(data,   
 sjPlot::sjt.xtab(Q57, Q260))