week_4

Diana Hilleshein

2/7/2022

Testing

```
library(tidyverse)
```

```
## -- 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   2.1.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
test <- tribble(
  ~A, ~B,  ~C,
  10, 10, 10,
  10, 10, NA,
  10, NA, NA,
  NA, NA, NA
)

View(test)

# Create a new variable which is sum of variables A, B and C
test$sum1 <- test$A + test$B + test$C #<= [1:1] + [1:2] + [1:3] and so on
test$sum1 #if there is at least one NA, the addition result will be NA
```

```
## [1] 30 NA NA NA
```

```
test$sum2 <- rowSums(test[,1:3]) #[rows,cols]//[all_rows,from 1st to 3rd cols]
test$sum2 #same...
```

```
## [1] 30 NA NA NA
```

```
test$sum2 <- rowSums(test[,1:3], na.rm = TRUE)
test$sum2 #na.rm helped to escape NA and get summ of the rows (NA as 0)
```

```
## [1] 30 20 10  0
```

```
test$mean1 <- (test$A + test$B + test$C)/3
test$mean1 #if there is at least one NA, the addition result will be NA
```

```
## [1] 10 NA NA NA
```

```
test$mean2 <- rowMeans(test[,1:3])
test$mean2 #same
```

```
## [1] 10 NA NA NA
```

```
test$mean2 <- rowMeans(test[,1:3], na.rm = TRUE)
test$mean2 #na.rm helped to escape NA and get mean of the rows (taking NA as 0 the
↪  function cant give a result since 0 cant cant be divided)
```

```
## [1]  10  10  10 NaN
```

Create a sum variable.

I will use the following variables:
Attitude to immigrants 4 levels 1) **imsmetn** -Allow many/few immigrants of same race/ethnic group as majority (scale from 1 to 4, 1 = Allow many to come and live here, 4 = Allow none)
2) **imdfetn** -Allow many/few immigrants of different race/ethnic group from majority (scale from 1 to 4, 1 = Allow many to come and live here, 4 = Allow none)
3) **impcntr** -Allow many/few immigrants from poorer countries outside Europe (scale from 1 to 4, 1 = Allow many to come and live here, 4 = Allow none)

```
setwd("C:/R scripts/DWA2022")
library(haven)

# import data
data <- read_sav("C:/R scripts/DWA2022/ESS9e03_1.sav")

save(data,file="data.csv")
head(data)
```

```
## # A tibble: 6 x 572
##   name   essround edition proddate  idno cntry    nwspol netusoft netustm ppltrst
##   <chr>     <dbl> <chr>   <chr>    <dbl> <chr+lb> <dbl+> <dbl+lb> <dbl+l> <dbl+l>
## 1 ESS9~         9 3.1     17.02.2~    27 AT [Aus~     60 5 [Ever~     180   2 [2]
## 2 ESS9~         9 3.1     17.02.2~   137 AT [Aus~     10 5 [Ever~      20   7 [7]
## 3 ESS9~         9 3.1     17.02.2~   194 AT [Aus~     60 4 [Most~     180   5 [5]
## 4 ESS9~         9 3.1     17.02.2~   208 AT [Aus~     45 5 [Ever~     120   3 [3]
## 5 ESS9~         9 3.1     17.02.2~   220 AT [Aus~     30 1 [Neve~      NA   5 [5]
## 6 ESS9~        9 3.1     17.02.2~   254 AT [Aus~     45 2 [Only~      NA   8 [8]
## # ... with 562 more variables: pplfair <dbl+lbl>, pplhlp <dbl+lbl>,
## #   polintr <dbl+lbl>, psppsgva <dbl+lbl>, actrolga <dbl+lbl>,
## #   psppipla <dbl+lbl>, cptppola <dbl+lbl>, trstprl <dbl+lbl>,
## #   trstlgl <dbl+lbl>, trstplc <dbl+lbl>, trstplt <dbl+lbl>, trstprt <dbl+lbl>,
## #   trstep <dbl+lbl>, trstun <dbl+lbl>, vote <dbl+lbl>, prtvtcat <dbl+lbl>,
## #   prtvtdbe <dbl+lbl>, prtvtdbg <dbl+lbl>, prtvtgch <dbl+lbl>,
## #   prtvtbcy <dbl+lbl>, prtvtecz <dbl+lbl>, prtvede1 <dbl+lbl>, ...
```

```r
dim(data)#everything is right
```

```
## [1] 49519   572
```

```r
table(data$imsmetn, useNA = "ifany")
```

```
## 
##     1     2     3     4  <NA> 
## 11898 21612  9474  3472  3063
```

```r
table(data$imdfetn, useNA = "ifany")
```

```
## 
##     1     2     3     4  <NA> 
##  7273 18722 13628  6781  3115
```

```r
table(data$impcntr, useNA = "ifany")
```

```
## 
##     1     2     3     4  <NA> 
##  6833 17738 14768  8561  1619
```

I found some documentation regarding useNA = “ifany” that I didnt know before: https://stat.ethz.ch/R-manual/R-devel/library/base/html/table.html. *“useNA controls if the table includes counts of NA values: the allowed values correspond to never (”no”), only if the count is positive (“ifany”) and even for zero counts (“always”)“*

```r
library(summarytools)
library(sjlabelled)

freq(data$imsmetn)
```

```
## Frequencies  
## data$imsmetn  
## Label: Allow many/few immigrants of same race/ethnic group as majority  
## Type: Numeric  
## 
##                 Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------- --------- -------------- --------- --------------
##           1   11898     25.61          25.61     24.03          24.03
##           2   21612     46.52           72.13     43.64           67.67
##           3    9474     20.39           92.53     19.13           86.80
##           4    3472      7.47          100.00      7.01           93.81
##        <NA>    3063                                6.19          100.00
##       Total   49519    100.00          100.00    100.00          100.00
```

```
freq(data$imdfetn)
```

```
## Frequencies
## data$imdfetn
## Label: Allow many/few immigrants of different race/ethnic group from majority
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------- --------- -------------- --------- --------------
##           1    7273     15.67          15.67     14.69          14.69
##           2   18722     40.35          56.02     37.81          52.50
##           3   13628     29.37          85.39     27.52          80.02
##           4    6781     14.61         100.00     13.69          93.71
##        <NA>    3115                               6.29         100.00
##       Total   49519    100.00         100.00    100.00         100.00
```

```
freq(data$impcntr)
```

```
## Frequencies
## data$impcntr
## Label: Allow many/few immigrants from poorer countries outside Europe
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------- --------- -------------- --------- --------------
##           1    6833     14.27          14.27     13.80          13.80
##           2   17738     37.03          51.30     35.82          49.62
##           3   14768     30.83          82.13     29.82          79.44
##           4    8561     17.87         100.00     17.29          96.73
##        <NA>    1619                               3.27         100.00
##       Total   49519    100.00         100.00    100.00         100.00
```

```
library(dplyr)
library(ggplot2)
library(scales)

#data %>% select(data, imsmetn, imdfetn, impcntr) %>%
#  summarytools::freq()
```

The last two lines didn‘ work for me, the follwoing error message appeared: *“Error: Must subset columns with a valid subscript vector.”* I faced this kind of mistke before and did not find a solution :(

So I decided to do it another way:

```r
#sel_var <- select(data, imsmetn, imdfetn, impcntr)
which(colnames(data) == "imsmetn")
```

```
## [1] 109
```

```r
which(colnames(data) == "imdfetn")
```

```
## [1] 110
```

```r
which(colnames(data) == "impcntr")
```

```
## [1] 111
```

```r
freq(data[,109:111])
```

```
## Frequencies
## data[, 109:111]$imsmetn
## Label: Allow many/few immigrants of same race/ethnic group as majority
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------- --------- -------------- --------- --------------
##           1   11898     25.61          25.61     24.03          24.03
##           2   21612     46.52          72.13     43.64          67.67
##           3    9474     20.39          92.53     19.13          86.80
##           4    3472      7.47         100.00      7.01          93.81
##        <NA>    3063                               6.19         100.00
##       Total   49519    100.00         100.00    100.00         100.00
##
## data[, 109:111]$imdfetn
## Label: Allow many/few immigrants of different race/ethnic group from majority
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------- --------- -------------- --------- --------------
##           1    7273     15.67          15.67     14.69          14.69
##           2   18722     40.35          56.02     37.81          52.50
##           3   13628     29.37          85.39     27.52          80.02
##           4    6781     14.61         100.00     13.69          93.71
##        <NA>    3115                               6.29         100.00
##       Total   49519    100.00         100.00    100.00         100.00
##
## data[, 109:111]$impcntr
## Label: Allow many/few immigrants from poorer countries outside Europe
## Type: Numeric
##
##                Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
## ----------- ------- --------- -------------- --------- --------------
##           1    6833     14.27          14.27     13.80          13.80
##           2   17738     37.03          51.30     35.82          49.62
##           3   14768     30.83          82.13     29.82          79.44
##           4    8561     17.87         100.00     17.29          96.73
##        <NA>    1619                               3.27         100.00
##       Total   49519    100.00         100.00    100.00         100.00
```

Well, for some reason this way worked! R can be unpredictable.

Next, I will create a sum variable.

```
data$sum <- rowSums(data[,109:111], na.rm = T)
head(data[,109:111], n = 25)
```

```
## # A tibble: 25 x 3
##                                  imsmetn         imdfetn         impcntr
##                                <dbl+lbl>       <dbl+lbl>       <dbl+lbl>
##   1 2 [Allow some]                       2 [Allow some]  2 [Allow some]
##   2 2 [Allow some]                       3 [Allow a few] 3 [Allow a few]
##   3 2 [Allow some]                       2 [Allow some]  3 [Allow a few]
##   4 2 [Allow some]                       3 [Allow a few] 3 [Allow a few]
##   5 3 [Allow a few]                      3 [Allow a few] 3 [Allow a few]
##   6 2 [Allow some]                       2 [Allow some]  2 [Allow some]
##   7 1 [Allow many to come and live here] 2 [Allow some]  2 [Allow some]
##   8 2 [Allow some]                       3 [Allow a few] 4 [Allow none]
##   9 4 [Allow none]                       4 [Allow none]  4 [Allow none]
##  10 2 [Allow some]                       2 [Allow some]  2 [Allow some]
## # ... with 15 more rows
```

```
head(data$sum, n = 25)
```

```
##  [1]  6  8  7  8  9  6  5  9 12  6  3  6  6 12  8  5  6  6  6 10  9  8 11  0  6
```

```
which(colnames(data) == "sum")
```

```
## [1] 573
```

The result from 24th row is 0, we can see that 24th row in the dataset has NA in all columns, therefore I get 0.

Find mean:

```
data$avg <- rowMeans(data[,109:111], na.rm = T)
head(data[,109:111], n = 25)
```

```
## # A tibble: 25 x 3
##                                  imsmetn         imdfetn         impcntr
##                                <dbl+lbl>       <dbl+lbl>       <dbl+lbl>
##   1 2 [Allow some]                       2 [Allow some]  2 [Allow some]
##   2 2 [Allow some]                       3 [Allow a few] 3 [Allow a few]
##   3 2 [Allow some]                       2 [Allow some]  3 [Allow a few]
##   4 2 [Allow some]                       3 [Allow a few] 3 [Allow a few]
##   5 3 [Allow a few]                      3 [Allow a few] 3 [Allow a few]
##  6 2 [Allow some]                       2 [Allow some]  2 [Allow some]
##  7 1 [Allow many to come and live here] 2 [Allow some]  2 [Allow some]
##  8 2 [Allow some]                       3 [Allow a few] 4 [Allow none]
##  9 4 [Allow none]                       4 [Allow none]  4 [Allow none]
## 10 2 [Allow some]                       2 [Allow some]  2 [Allow some]
## # ... with 15 more rows
```

```
head(data$avg, n = 25)
```

```
##  [1] 2.000000 2.666667 2.333333 2.666667 3.000000 2.000000 1.666667 3.000000
##  [9] 4.000000 2.000000 1.000000 2.000000 2.000000 4.000000 2.666667 1.666667
## [17] 2.000000 2.000000 2.000000 3.333333 3.000000 2.666667 3.666667      NaN
## [25] 2.000000
```

```
which(colnames(data) == "avg")
```

```
## [1] 574
```

The 24th row does not have a result since 0 can not be devided.

Incorrect versions:

```
data$wrong_sum <- data$imsmetn + data$imdfetn + data$impcntr
head(data$wrong_sum, n = 25)
```

```
##  [1]  6  8  7  8  9  6  5  9 12  6  3  6  6 12  8  5  6  6  6 10  9  8 11 NA  6
```

```
data$wrong_avg <- (data$imsmetn + data$imdfetn + data$impcntr)/3
head(data$wrong_avg, n = 25)
```

```
##  [1] 2.000000 2.666667 2.333333 2.666667 3.000000 2.000000 1.666667 3.000000
##  [9] 4.000000 2.000000 1.000000 2.000000 2.000000 4.000000 2.666667 1.666667
## [17] 2.000000 2.000000 2.000000 3.333333 3.000000 2.666667 3.666667       NA
## [25] 2.000000
```

Same situation with 24th row.

Calculate some descriptive statistics of your new sum variable and visualize it.

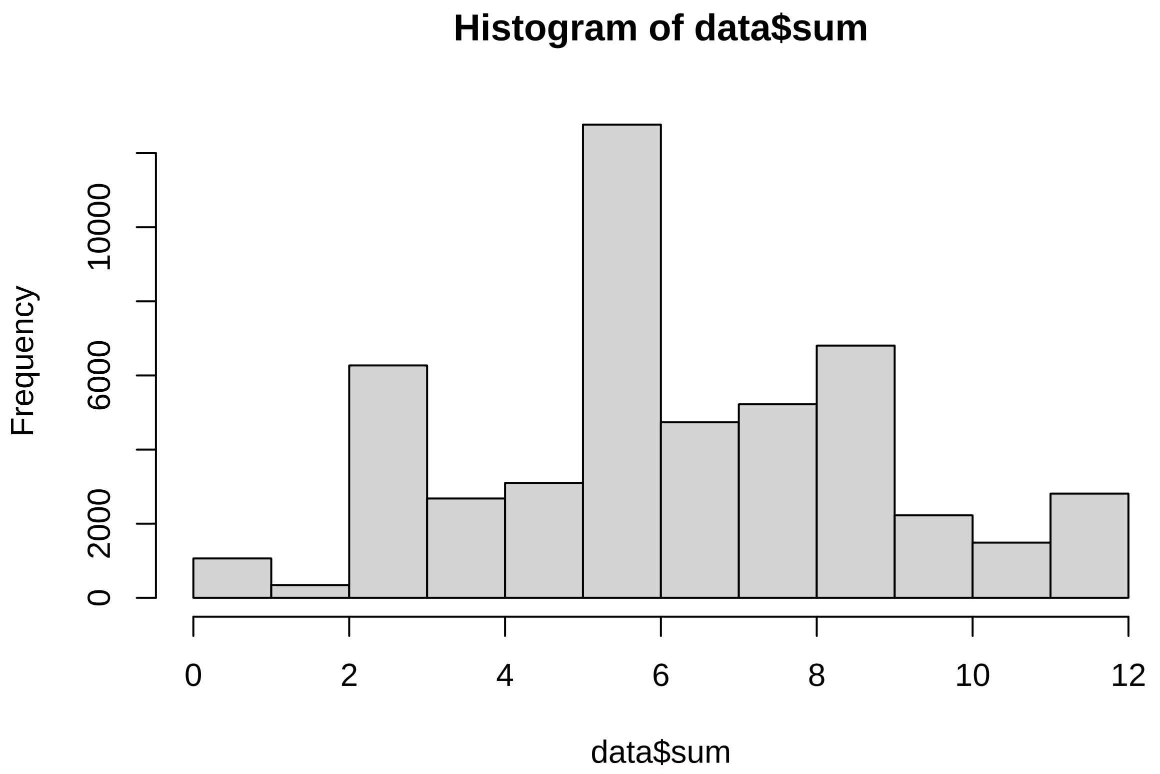
```
summary(data[,109:111])
```

```
##     imsmetn         imdfetn         impcntr
##  Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :2.000
##  Mean   :2.097   Mean   :2.429   Mean   :2.523
##  3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:3.000
##  Max.   :4.000   Max.   :4.000   Max.   :4.000
##  NA's   :3063    NA's   :3115    NA's   :1619
```

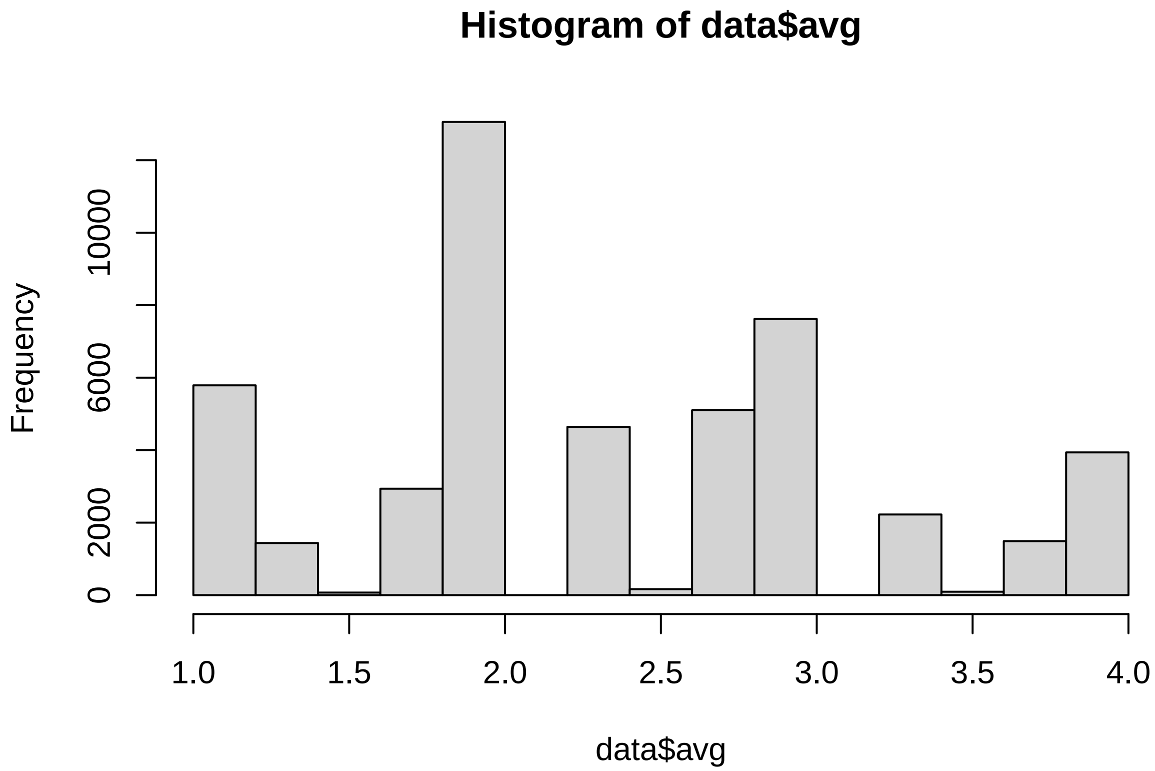
```
summary(data[,573:574])
```

```
##       sum              avg       
##  Min.   : 0.000   Min.   :1.000  
##  1st Qu.: 5.000   1st Qu.:2.000  
##  Median : 6.000   Median :2.333  
##  Mean   : 6.685   Mean   :2.378  
##  3rd Qu.: 9.000   3rd Qu.:3.000  
##  Max.   :12.000   Max.   :4.000  
##                   NA's   :956
```

```
hist(data$sum)
```



```
hist(data$avg)
```



The majority of answers summarized fall on number 5 that I can interpret as above average. The majority of answers in plotted for average values fall on 2, that is again above average.

Use some sub-groups or domains in the data set (e.g. country, gender, agegroup, . . .) and calculate some descriptive statistics of your new sum variable and visualize it by these sub-groups/domains.

Before I start it is important to note that sum variable has values from 0 to 12 - from “Allow many to come” to “Allow none”,

```
w4 <-select(data, cntry, agea, gndr, sum)
w4
```

```
## # A tibble: 49,519 x 4
##    cntry             agea       gndr   sum
##    <chr+lbl>    <dbl+lbl>  <dbl+lbl> <dbl>
##  1 AT [Austria]        43 1 [Male]       6
##  2 AT [Austria]        67 1 [Male]       8
##  3 AT [Austria]        40 2 [Female]     7
##  4 AT [Austria]        63 1 [Male]       8
##  5 AT [Austria]        71 2 [Female]     9
##  6 AT [Austria]        64 1 [Male]       6
##  7 AT [Austria]        56 1 [Male]       5
##  8 AT [Austria]        74 2 [Female]     9
##  9 AT [Austria]        37 1 [Male]      12
## 10 AT [Austria]        22 2 [Female]     6
## # ... with 49,509 more rows
```

```
females <- w4 %>% filter(gndr == 2)
males <- w4 %>% filter(gndr == 1)

summary(females)
```

```
##     cntry                agea            gndr        sum
##  Length:26499       Min.   :15.00   Min.   :2   Min.   : 0.00
##  Class :character   1st Qu.:37.00   1st Qu.:2   1st Qu.: 5.00
##  Mode  :character   Median :53.00   Median :2   Median : 6.00
##                     Mean   :51.68   Mean   :2   Mean   : 6.64
##                     3rd Qu.:67.00   3rd Qu.:2   3rd Qu.: 9.00
##                     Max.   :90.00   Max.   :2   Max.   :12.00
##                     NA's   :110
```

```
summary(males)
```

```
##     cntry                agea            gndr        sum
##  Length:23020       Min.   :15.00   Min.   :1   Min.   : 0.000
##  Class :character   1st Qu.:36.00   1st Qu.:1   1st Qu.: 5.000
##  Mode  :character   Median :51.00   Median :1   Median : 6.000
##                     Mean   :50.35   Mean   :1   Mean   : 6.736
##                     3rd Qu.:65.00   3rd Qu.:1   3rd Qu.: 9.000
##                     Max.   :90.00   Max.   :1   Max.   :12.000
##                     NA's   :112
```

The difference in sum value is not that big between men and women, therefore there is no a big difference between an attitude to immigrants between females and males.

```
mean(w4$agea, na.rm = T) #51 years is an average age
```

```
## [1] 51.06601
```

```
younger <- w4 %>% filter(agea < 51)
older <- w4 %>% filter(agea >= 51)

summary(younger)
```

```
##     cntry                agea            gndr            sum
##  Length:23460       Min.   :15.00   Min.   :1.000   Min.   : 0.000
##  Class :character   1st Qu.:26.00   1st Qu.:1.000   1st Qu.: 5.000
##  Mode  :character   Median :35.00   Median :2.000   Median : 6.000
##                     Mean   :34.52   Mean   :1.523   Mean   : 6.391
##                     3rd Qu.:43.00   3rd Qu.:2.000   3rd Qu.: 8.000
##                     Max.   :50.00   Max.   :2.000   Max.   :12.000
```

```
summary(older)
```

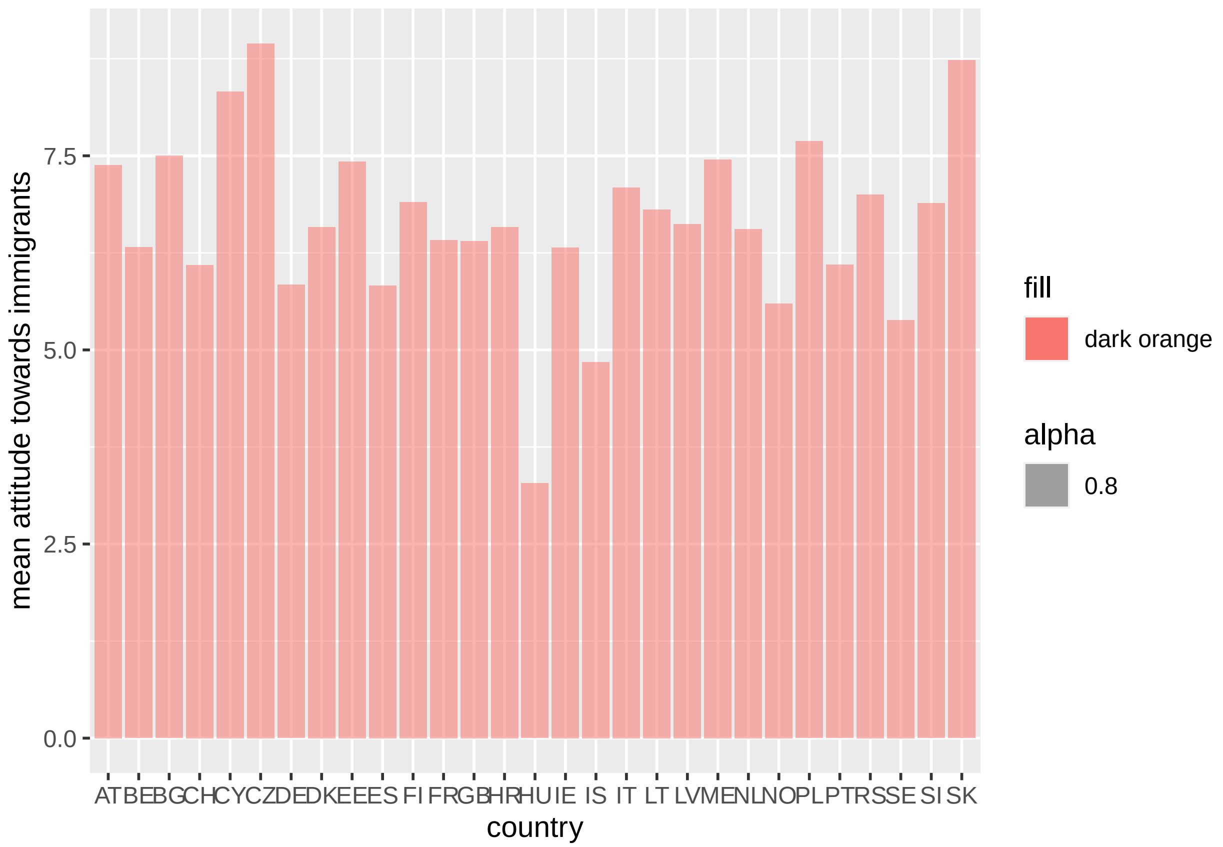
```
##      cntry                agea            gndr            sum
##  Length:25837       Min.   :51.00   Min.   :1.000   Min.   : 0.000
##  Class :character   1st Qu.:58.00   1st Qu.:1.000   1st Qu.: 6.000
##  Mode  :character   Median :65.00   Median :2.000   Median : 7.000
##                     Mean   :66.09   Mean   :1.547   Mean   : 6.955
##                     3rd Qu.:73.00   3rd Qu.:2.000   3rd Qu.: 9.000
##                     Max.   :90.00   Max.   :2.000   Max.   :12.000
```

There is no a significant difference between an attitude to immigrants between people who are younger or older that 51.

```
library(dplyr)
a1 <- w4 %>% group_by(cntry) %>% summarise(mean_sum = mean(sum, na.rm = TRUE))
a1
```

```
## # A tibble: 29 x 2
##    cntry             mean_sum
##    <chr+lbl>            <dbl>
##  1 AT [Austria]          7.38
##  2 BE [Belgium]          6.32
##  3 BG [Bulgaria]         7.50
##  4 CH [Switzerland]      6.09
##  5 CY [Cyprus]           8.33
##  6 CZ [Czechia]          8.95
##  7 DE [Germany]          5.84
##  8 DK [Denmark]          6.58
##  9 EE [Estonia]          7.43
## 10 ES [Spain]            5.83
## # ... with 19 more rows
```

```
ggplot(a1, aes(x = cntry, y = mean_sum, fill = "dark orange", alpha = 0.8)) +
  geom_bar(stat="identity", position=position_dodge()) + ylab ("mean attitude towards
  ↪  immigrants") + xlab("country")
```



* AT = Austria, BE = Belgium, BG = Bulgaria, CH = Switzerland, CY = Cyprus, CZ = Czechia, DE = Germany, DK = Denmark, EE = Estonia, ES = Spain, FI = Finland, FR = France, GB = United Kingdom, HR = Croatia, HU = Hungary, IE = Ireland, IS = Iceland, IT = Italy, LT = Lithuania, LV = Latvia, ME = Montenegro, NL = Netherlands, NO = Norway, PL = Poland, PT = Portugal, RS = Serbia, SE = Sweden, SI = Slovenia, SK = Slovakia

The interpretation for results: 0 - Allow many to come and live here,… 12 - Allow none

We can see that people in Hungary and Iceland have the lowest value (below 5) for mean attitude towards immigrants. I feel that results for hungaria may be unreliable knowing politics of Hungarian government. Majority of values fall between 5 and 7.