week 3

Diana Hilleshein

2/2/2022

```
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>, ...
```

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

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

Part 1. Subsetting (rows)

Create subfiles (e.g. use gender or country or…)

```
library(tidyverse)
library(sjlabelled)
get_labels(data$gndr) #to know labels for the categories
```

```
## [1] "Male"      "Female"    "No answer"
```

```
table(data$gndr) # there are 23020 males and 26499 females
```

```
##
##     1     2
## 23020 26499
```

```
# sub for males
males <- data %>% filter(gndr == 1) # 1 for males
table(males$gndr) #number in the table matches number of males that we got before
```

```
##
##     1
## 23020
```

```
dim(males) # double check. Again number of observations is the same as number of males,
↪  so we successfully made the subsetting for males
```

```
## [1] 23020   572
```

Subsetting for males is done, now I ell do the same for females.

```
females <- data %>% filter(gndr == 2) # 2 for females
table(females$gndr) # 26499
```

```
##
##     2
## 26499
```

```
dim(females)
```

```
## [1] 26499   572
```

Subsetting for females is done. Both subsets have right number of observations and 572 cases.

Try to combine subfiles. Report your attempts.

```
# Combine datasets
comb_data <- rbind(males, females) # df1 = males, df2 = females
table(comb_data$gndr)
```

```
##
##     1     2
## 23020 26499
```

To understand the “rbind” function I googled how to use it: https://www.statology.org/rbind-in-r/, in the “Example 4: Rbind Two Data Frames” section we can see an example.
After I combines the subsets, I checked the number of males and females in the combined set. The number of males and females is right!

Part 2. Merging

Create two files where you select some variables. Remember to include variables CNTRY and IDNO into both of the files. Those will be your "key" variables when you merge/combine these two datasets.

```
#create two sets
set1 <- select(data, idno, cntry, gndr)
set2 <- select(data, idno, cntry, stflife)
head(set1) #all good
```

```
## # A tibble: 6 x 3
##    idno cntry                gndr
##   <dbl> <chr+lbl>     <dbl+lbl>
## 1    27 AT [Austria] 1 [Male]
## 2   137 AT [Austria] 1 [Male]
## 3   194 AT [Austria] 2 [Female]
## 4   208 AT [Austria] 1 [Male]
## 5   220 AT [Austria] 2 [Female]
## 6   254 AT [Austria] 1 [Male]
```

```
head(set2) #all good
```

```
## # A tibble: 6 x 3
##    idno cntry            stflife
##   <dbl> <chr+lbl>    <dbl+lbl>
## 1    27 AT [Austria]     8 [8]
## 2   137 AT [Austria]     8 [8]
## 3   194 AT [Austria]     9 [9]
## 4   208 AT [Austria]     8 [8]
## 5   220 AT [Austria]     8 [8]
## 6   254 AT [Austria]     9 [9]
```

Try to merge files. Remember to use CNTRY and IDNO variables as keys. Report your attempts.

```
# Merge
full_set <- merge(set1, set2, by = c("cntry", "idno"))
View(full_set) #success!
head(full_set)
```

```
##   cntry  idno gndr stflife
## 1    AT 10008    1       8
## 2    AT  1005    1       9
## 3    AT 10062    1       8
## 4    AT 10075    1       9
## 5    AT 10082    2       4
## 6    AT 10097    1       2
```

Indeed, examining the data i could see that ID numbers for participants were the same, while countries were different. For example:
ID cntry gndr
2 BG 2
2 CZ 1

Ir was easy to make two files with different variable and merge them using idno and cntry as key variables. I didn't have any problems since I have already done it before. It was long time ago, so, some practice was nice to refresh my skills.

Part 3. Aggregating

For the task I will use following variables:
1) **gndr**-Gender (1 = Male, 2 = Female)
2) **agea** -Age of respondent, calculated (integer)
3) **cntry** - Country of respondent (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)
4) **stflife** -How satisfied with life as a whole (from 0 to 10, 0 = extremely dissatisfied, 10 = extremely satisfied)

Aggregate by gender and calculate some descriptives for one or some variable(s)

I didn't know what is na.rm and decided to Google. This website helped me a lot to understand it: https://www.programmingr.com/tutorial/na-rm/.

```
table(data$agea)
```

```
##
##  15  16  17  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34
## 247 445 469 507 599 494 533 534 543 548 480 612 601 610 615 654 637 658 657 655
##  35  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53  54
## 693 701 730 701 813 800 747 747 763 737 777 801 749 858 872 873 774 886 897 918
##  55  56  57  58  59  60  61  62  63  64  65  66  67  68  69  70  71  72  73  74
## 914 861 860 915 872 868 864 872 946 841 836 912 861 860 877 843 804 757 675 593
##  75  76  77  78  79  80  81  82  83  84  85  86  87  88  89  90
## 554 504 564 532 540 474 354 346 293 271 205 226 158 124 120 266
```

```
table(data$gndr)
```

```
##
##     1     2
## 23020 26499
```

```
table(data$cntry)
```

```
##
##    AT    BE    BG    CH    CY    CZ    DE    DK    EE    ES    FI    FR    GB    HR    HU    IE
## 2499 1767 2198 1542  781 2398 2358 1572 1904 1668 1755 2010 2204 1810 1661 2216
##   IS   IT   LT   LV   ME   NL   NO   PL   PT   RS   SE   SI   SK
##  861 2745 1835  918 1200 1673 1406 1500 1055 2043 1539 1318 1083
```

```
table(data$stflife)
```

```
##
##     0     1     2     3     4     5     6     7     8     9    10
##   720   455   968  1623  1935  5005  4246  8490 12700  7384  5729
```

```
# aggregate by gender. Note: na.rm = TRUE
a1 <- data %>% group_by(gndr) %>%
  summarise(n_stflife = n(), mean_stflife = mean(stflife, na.rm = TRUE)) #men are a bit
  ↪  more satisfied with life then women. But in average level of satisfaction is above
  ↪  average for man and for women.
a1
```

```
## # A tibble: 2 x 3
##         gndr n_stflife mean_stflife
##    <dbl+lbl>     <int>        <dbl>
## 1 1 [Male]       23020         7.15
## 2 2 [Female]     26499         7.08
```

Aggregate by country and calculate some descriptives for one or some variable(s)

```
a2 <- data %>% group_by(gndr, cntry) %>%
  summarise(n_stflife = n(), mean_stflife = mean(stflife, na.rm = TRUE))
head(a2) #print just first 2 pages to not make the report too large
```

```
## # A tibble: 6 x 4
## # Groups:   gndr [1]
##        gndr cntry            n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>            <int>        <dbl>
## 1  1 [Male] AT [Austria]          1153         7.78
## 2  1 [Male] BE [Belgium]           868         7.56
## 3  1 [Male] BG [Bulgaria]          976         4.96
## 4  1 [Male] CH [Switzerland]       775         8.13
## 5  1 [Male] CY [Cyprus]            366         7.11
## 6  1 [Male] CZ [Czechia]          1049         6.97
```

```
a2[which(a2$mean_stflife == max(a2$mean_stflife)), ] #women in Denmark are the most
↪  satisfied with life
```

```
## # A tibble: 1 x 4
## # Groups:   gndr [1]
##         gndr cntry        n_stflife mean_stflife
##    <dbl+lbl> <chr+lbl>        <int>        <dbl>
## 1 2 [Female] DK [Denmark]       726         8.54
```

```
a2[which(a2$mean_stflife == min(a2$mean_stflife)), ] #women in Bulgaria are the lest
↪  satisfied with life
```

```
## # A tibble: 1 x 4
## # Groups:   gndr [1]
##         gndr cntry         n_stflife mean_stflife
##    <dbl+lbl> <chr+lbl>         <int>        <dbl>
## 1 2 [Female] BG [Bulgaria]      1222         4.89
```

```
# What happens if we don`t separate by gender?
a3 <- data %>% group_by(cntry) %>%
  summarise(n_stflife = n(), mean_stflife = mean(stflife, na.rm = TRUE))
head(a3)
```

```
## # A tibble: 6 x 3
##   cntry            n_stflife mean_stflife
##   <chr+lbl>            <int>        <dbl>
## 1 AT [Austria]          2499         7.85
## 2 BE [Belgium]          1767         7.53
## 3 BG [Bulgaria]         2198         4.92
## 4 CH [Switzerland]      1542         8.15
## 5 CY [Cyprus]            781         7.02
## 6 CZ [Czechia]          2398         7.02
```

```
a3[which(a3$mean_stflife == max(a3$mean_stflife)), ] #people in Denmark are still the
↪  most satisfied with life
```

```
## # A tibble: 1 x 3
##   cntry        n_stflife mean_stflife
##   <chr+lbl>        <int>        <dbl>
## 1 DK [Denmark]      1572         8.50
```

```
a3[which(a3$mean_stflife == min(a3$mean_stflife)), ] #people in Bulgaria are still the
↪  lest satisfied with life
```

```
## # A tibble: 1 x 3
##   cntry         n_stflife mean_stflife
##   <chr+lbl>         <int>        <dbl>
## 1 BG [Bulgaria]      2198         4.92
```

Therefore, it seems like women have wider range of satisfaction/unsatisfactory, but generally speaking the results of women are closer to results of men.

```
# What is the most and the least satisfied age?
a4 <- data %>% group_by(agea) %>%
  summarise(n_stflife = n(), mean_stflife = mean(stflife, na.rm = TRUE))
head(a4)
```

```
## # A tibble: 6 x 3
##        agea n_stflife mean_stflife
##   <dbl+lbl>     <int>        <dbl>
## 1        15       247         8.03
## 2        16       445         8.15
## 3        17       469         7.90
## 4        18       507         7.80
## 5        19       599         7.55
## 6        20       494         7.59
```

```
a4[which(a4$mean_stflife == max(a4$mean_stflife)), ] #people aged 16 tend to be the most
↪  satisfied with life
```

```
## # A tibble: 1 x 3
##        agea n_stflife mean_stflife
##   <dbl+lbl>     <int>        <dbl>
## 1        16       445         8.15
```

```
a4[which(a4$mean_stflife == min(a4$mean_stflife)), ] #people aged 83 tend to be the lest
↪  satisfied with life, but, happily, still above average!
```

```
## # A tibble: 1 x 3
##        agea n_stflife mean_stflife
##   <dbl+lbl>     <int>        <dbl>
## 1        83       293         6.62
```

```
#Now I want to see difference by age and countries
a5 <- data %>% group_by(agea, cntry) %>%
  summarise(n_stflife = n(), mean_stflife = mean(stflife, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'agea'. You can override using the `.groups` argument.
```

```
head(a5)
```

```
## # A tibble: 6 x 4
## # Groups:   agea [1]
##        agea cntry             n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>             <int>        <dbl>
## 1        15 AT [Austria]              6         8.17
## 2        15 BE [Belgium]             12         7.42
## 3        15 BG [Bulgaria]            24         7.52
## 4        15 CH [Switzerland]         11         8.45
## 5        15 CY [Cyprus]               2         8
## 6        15 CZ [Czechia]              5         8.2
```

```
a5[which(a5$mean_stflife == max(a5$mean_stflife)), ]
```

```
## # A tibble: 2 x 4
## # Groups:   agea [1]
##        agea cntry        n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>        <int>        <dbl>
## 1        88 DK [Denmark]         4           10
## 2        88 IS [Iceland]         1           10
```

```
a5[which(a5$mean_stflife == min(a5$mean_stflife)), ] #this code gave NA in age, so I need
↪  to check if there are missing values. Nevertheless, there is a person from Bulgaria
↪  of unknown age who feels very unsatisfied with life ):
```

```
## # A tibble: 1 x 4
## # Groups:   agea [1]
##        agea cntry         n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>         <int>        <dbl>
## 1        NA BG [Bulgaria]         1            1
```

```
library(tidyr)
sum(is.na(data$agea)) # there are 222 missing values, so I need to change the code. I
↪  didn`t find a good way to show if there are NA...
```

```
## [1] 222
```

```
data <- data %>% drop_na(agea)

a5 <- data %>% group_by(agea, cntry) %>%
  summarise(n_stflife = n(), mean_stflife = mean(stflife, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'agea'. You can override using the `.groups` argument.
```

```
head(a5)
```

```
## # A tibble: 6 x 4
## # Groups:   agea [1]
##        agea cntry            n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>            <int>        <dbl>
## 1        15 AT [Austria]             6         8.17
## 2        15 BE [Belgium]            12         7.42
## 3        15 BG [Bulgaria]           24         7.52
## 4        15 CH [Switzerland]        11         8.45
## 5        15 CY [Cyprus]              2         8
## 6        15 CZ [Czechia]             5         8.2
```

```
a5[which(a5$mean_stflife == max(a5$mean_stflife)), ] #people from Denmark are still
↪  leading in being the most satisfied! People aged 88 from Denmark and Island are the
↪  most satisfied.
```

```
## # A tibble: 2 x 4
## # Groups:   agea [1]
##        agea cntry        n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>        <int>        <dbl>
## 1        88 DK [Denmark]         4           10
## 2        88 IS [Iceland]         1           10
```

```
a5[which(a5$mean_stflife == min(a5$mean_stflife)), ] #people from Portugal aged 90 are
↪  the least satisfied.
```

```
## # A tibble: 1 x 4
## # Groups:   agea [1]
##        agea cntry         n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>         <int>        <dbl>
## 1        90 PT [Portugal]         3         2.33
```

I looked at extreme (max/min) indicators of mean_lifestyle. Nevertheless, it would be right to see distributions for the findings.

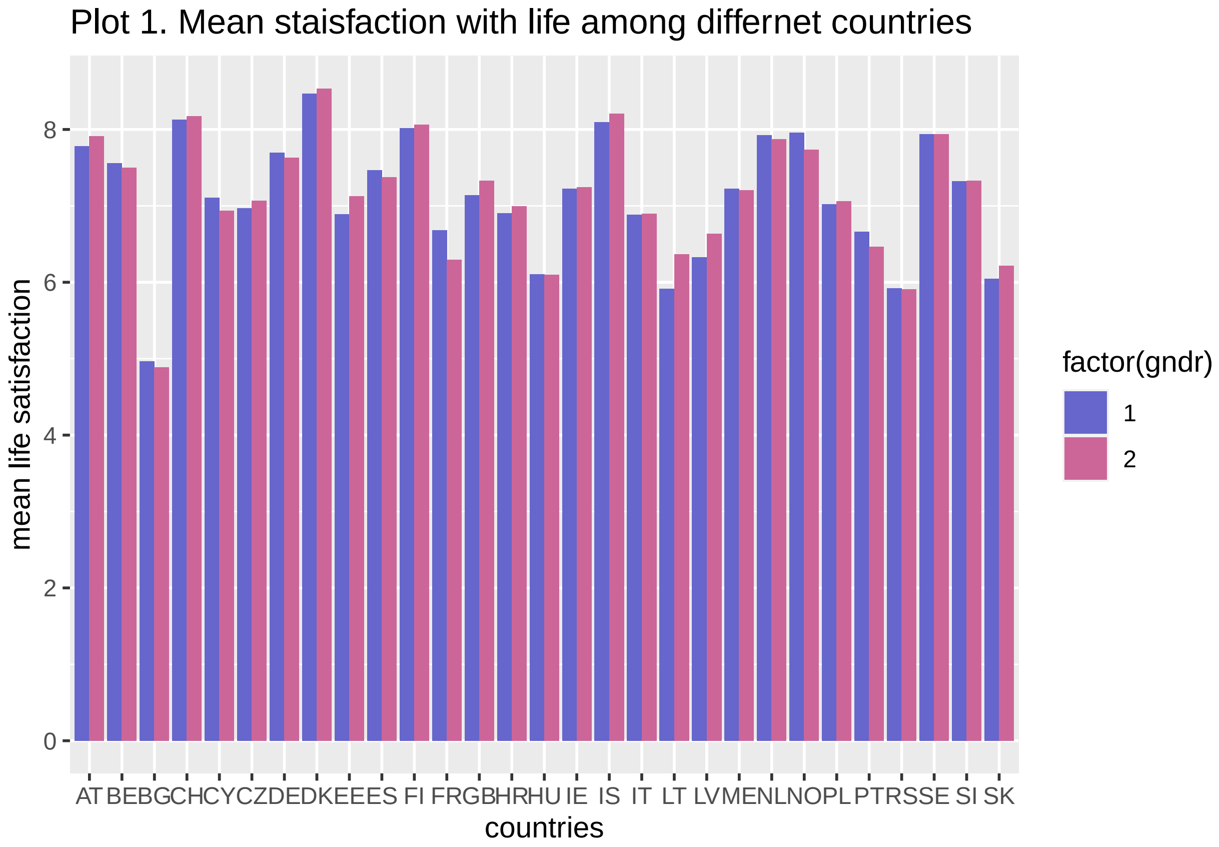
```
options(tibble.print_max = Inf)
# I will be interested in a2 and a5 aggregations

head(a2)
```

```
## # A tibble: 6 x 4
## # Groups:   gndr [1]
##        gndr cntry            n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>            <int>        <dbl>
## 1  1 [Male] AT [Austria]          1153         7.78
## 2  1 [Male] BE [Belgium]           868         7.56
## 3  1 [Male] BG [Bulgaria]          976         4.96
## 4  1 [Male] CH [Switzerland]       775         8.13
## 5  1 [Male] CY [Cyprus]            366         7.11
## 6  1 [Male] CZ [Czechia]          1049         6.97
```

```
head(a5)
```

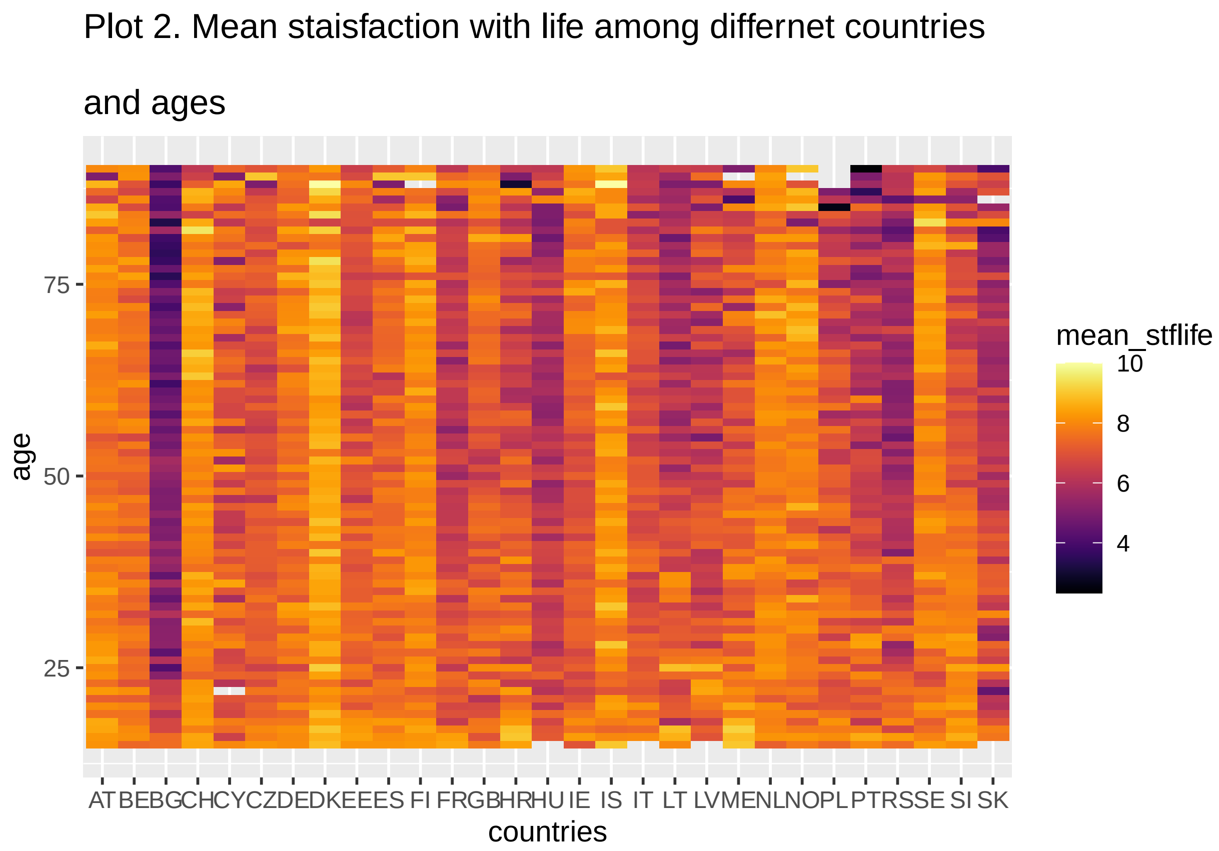
```
## # A tibble: 6 x 4
## # Groups:   agea [1]
##        agea cntry            n_stflife mean_stflife
##   <dbl+lbl> <chr+lbl>            <int>        <dbl>
## 1        15 AT [Austria]             6         8.17
## 2        15 BE [Belgium]            12         7.42
## 3        15 BG [Bulgaria]           24         7.52
## 4        15 CH [Switzerland]        11         8.45
## 5        15 CY [Cyprus]              2         8
## 6        15 CZ [Czechia]             5         8.2
```

```
#plotting
library(ggplot2)
ggplot(a2, aes(x = cntry, y = mean_stflife, fill = factor(gndr))) +
  geom_bar(stat="identity", position=position_dodge()) + scale_fill_manual(values = c("1"
  ↪  = "#6666CC",
  "2" = "#CC6699")) + xlab("countries") +  ylab("mean life satisfaction") + ggtitle("Plot
  ↪  1. Mean staisfaction with life among differnet countries")
```


Plot 1. Mean staisfaction with life among differnet countries

As I already have found, people from Bulgaria are the least satisfied in comparison to all countries. Serbia is on the second place of the least satisfied.
Meanwhile, people from Denmark are the happiest.

```
library(scales)
library(viridis)
ggplot(a5, aes(x = cntry, y = agea)) + geom_raster(aes(fill = mean_stflife))+
↪  scale_fill_gradient2(low="#FF9900") + scale_fill_viridis(option="B") +
↪  xlab("countries") +  ylab("age") + ggtitle("Plot 2. Mean staisfaction with life among
↪  differnet countries

↪  \nand ages")
```



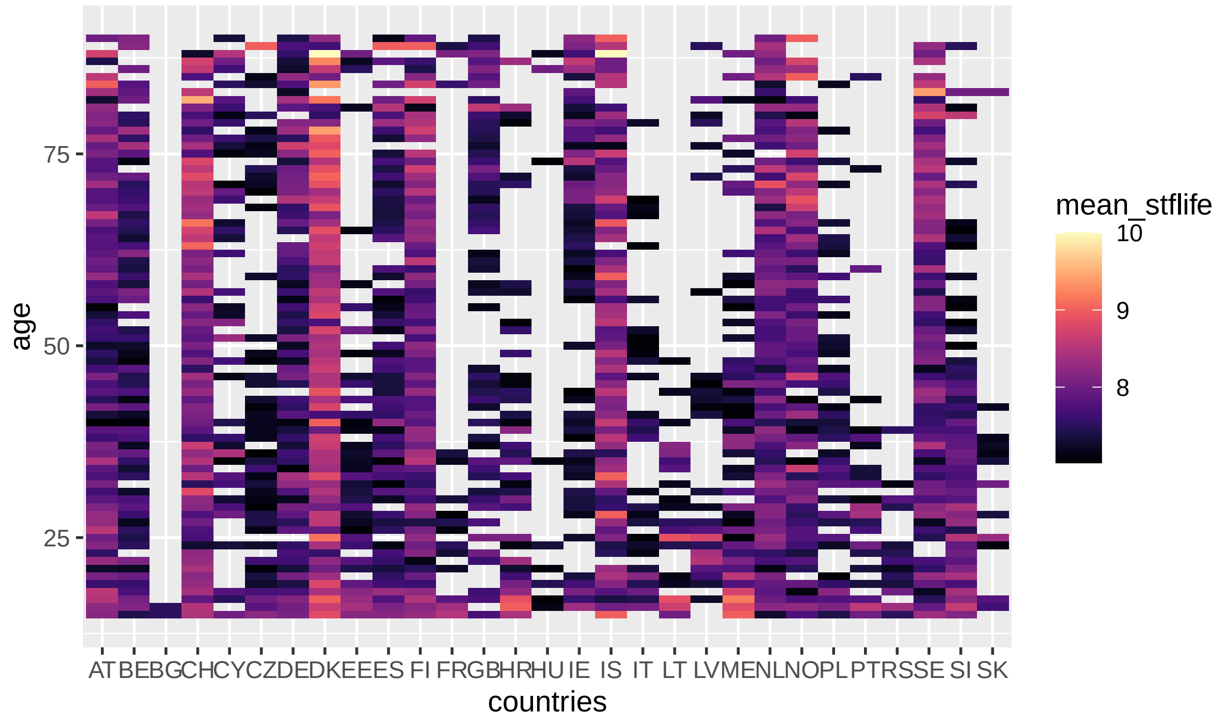
```
# for better vizualization
mean(a5$mean_stflife)
```

```
## [1] 7.160721
```

```
a5.1 <- a5 %>% filter(mean_stflife <= 7)
a5.2 <- a5 %>% filter(mean_stflife  > 7)



ggplot(a5.2, aes(x = cntry, y = agea)) + geom_raster(aes(fill = mean_stflife))+
↪  scale_fill_gradient2(low = "#ddd85e",
    high = "#6f1873",
    breaks = pretty_breaks(n = 5)) + scale_fill_viridis(option="A")+ xlab("countries") +
    ↪  ylab("age") + ggtitle("Plot 3. Mean staisfaction with life among differnet
    ↪  countries

↪  \nand ages (above average)")
```

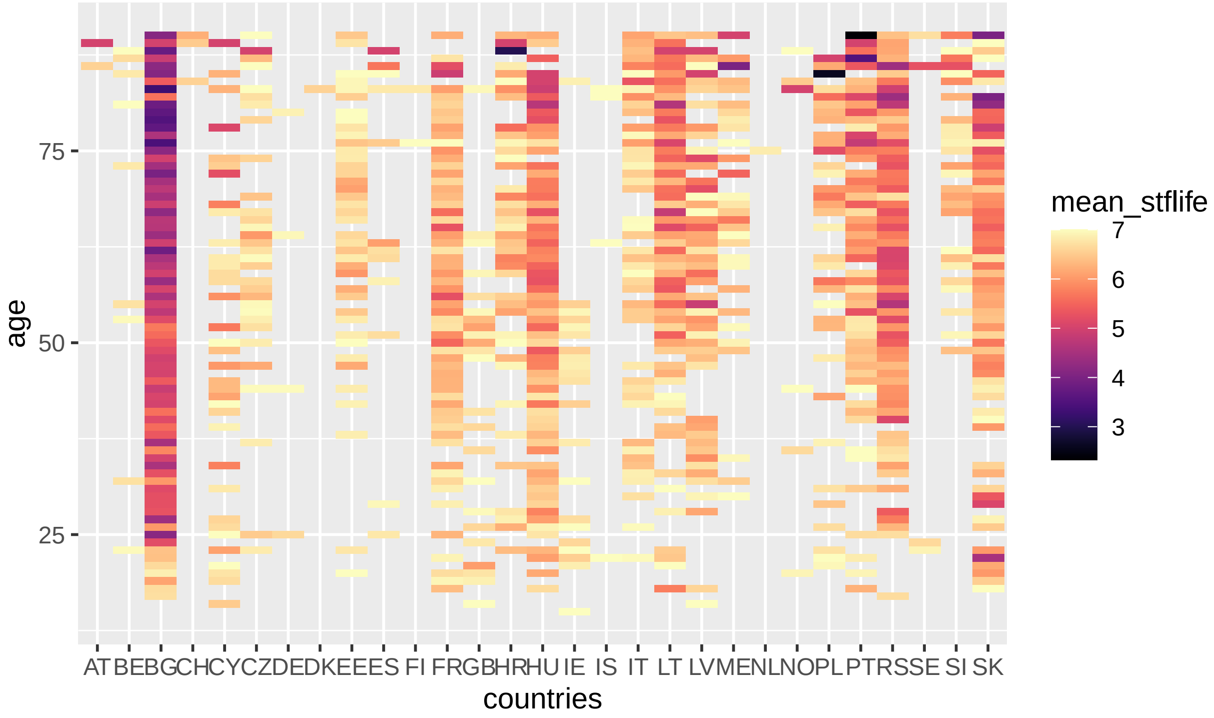
Plot 3. Mean staisfaction with life among differnet countries

and ages (above average)



```
ggplot(a5.1, aes(x = cntry, y = agea)) + geom_raster(aes(fill = mean_stflife))+
↪   scale_fill_gradient2(low = "#fbece3",
    high = "#d85edd",
    breaks = pretty_breaks(n = 5)) + scale_fill_viridis(option="A") + xlab("countries") +
    ↪   ylab("age") + ggtitle("Plot 4. Mean staisfaction with life among differnet
    ↪   countries

↪   \nand ages (below average)")
```

Plot 4. Mean staisfaction with life among differnet countries and ages (below average)



In average people are the most satisfied before 25 or after 70. In Nordic countries people who are over 70 are significantly more satisfied in comparison to other countries. An interesting finding is that people in Nordic countries tent to be more satisfied after 70 years rather then before 25. In countries such as Bulgaria, Portugal, Serbia people are happier at before they reach 25 and not significantly satisfied over 70. This difference is a pattern and is most likely related to economic conditions and social support in those countries. Developed and rich countries, such as Denmark, Norway, Austria have significantly higher level of overall satisfaction of people at any ages.

For more comfortable view, I separated the main raster plot in a way that the third would show data for people whose satisfaction with life above average (mean = 7.160721). The plot seems to be the most dense in the lover part indicating that people below 25 tent to be the most satisfied.
We can see that Austria, Denmark, Switzerland, Finland, Island, Norway, Sweden and Netherlands have the most light or the most filled lines that means that the lest number of people feeling above average satisfaction. Denmark, Switzerland and Island have notably higher average satisfaction.

The forth plot shows people dense satisfaction with life is below average. Bulgaria, Hungary, Portugal, France, Slovakia and Serbia have the majority of people whose satisfaction with life is below average.