Real data example

Giovanni Saraceno

Contents

```
library(tidyverse)
```

```
## Warning: il pacchetto 'tidyverse' è stato creato con R versione 4.3.2

## Warning: il pacchetto 'ggplot2' è stato creato con R versione 4.3.3

## Warning: il pacchetto 'stringr' è stato creato con R versione 4.3.3

## Warning: il pacchetto 'lubridate' è stato creato con R versione 4.3.2
```

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

The data comes from the study of Woodworth et al. (2017), a replication study of Seligman (2005)‘s work which had suggested that positive psychology interventions, when delivered via the internet, could increase participants' happiness and decrease their depression relative to the changes effected by a placebo control.

Their main finding was contrary to that of the original study by Seligman (2005). All interventions, including the theoretically-neutral placebo, led to significant increases in happiness and to significant reductions in depression. The effects of the positive-psychology interventions were statistically indistinguishable from those of the placebo.

We have two data sets:

- The first data set (`ahi-cesd.csv`) comprises 992 point-in-time records of the self-reported happiness and depression of 295 participants, each assigned to one of four intervention groups, in a study of the effect of web-based positive-psychology interventions on happiness and depression. Each point-in-time measurement consists of a participant's responses to the 24 items of the **Authentic Happiness Inventory (AHI)** and to the 20 items of the **Center for Epidemiological Studies Depression (CES-D)** scale. Measurements were attempted at the time of each participant's enrollment in the study and on 5 subsequent occasions, the last being approximately 189 days after enrollment.

A *total AHI score* is obtained by summing the scores for the 24 items. A *total CES-D score* is obtained by first reversing the scores of items 4, 8, 12 and 16, and then summing the scores for the 20 items.

The `ahi-cesd.csv` contains the following variables:

- `id`: Participant ID.
- `occasion`: Measurement occasion: 0 = Pretest (i.e., at enrollment), 1 = Posttest (i.e., 7 days after pretest), 2 = 1-week follow-up, (i.e., 14 days after pretest, 7 days after posttest), 3 = 1-month follow-up, (i.e., 38 days after pretest, 31 days after posttest), 4 = 3-month follow-up, (i.e., 98 days after pretest, 91 days after posttest), 5 = 6-month follow-up, (i.e., 189 days after pretest, 182 days after posttest).
- `elapsed.days`: Time since enrollment measured in fractional days.
- `intervention`: 3 positive psychology interventions (PPIs), plus 1 control condition: 1 = Using signature strengths, 2 = Three good things, 3 = Gratitude visit, 4 = Recording early memories (control condition).
- `ahi01-ahi24`: Responses on 24 AHI items.
- `cesd01-cesd20`: Responses on 20 CES-D items.

- `ahiTotal`: Total AHI score.
- `cesdTotal`: Total CES-D score.

```r
dat <- read.csv("ahi-cesd.csv")
str(dat)
```

```
## 'data.frame':    992 obs. of  50 variables:
##  $ id          : int  1 1 2 2 2 2 2 2 3 3 ...
##  $ occasion    : int  0 1 0 1 2 3 4 5 0 2 ...
##  $ elapsed.days: num  0 11.77 0 8.02 14.3 ...
##  $ intervention: int  4 4 1 1 1 1 1 1 4 4 ...
##  $ ahi01       : int  2 3 3 3 3 3 3 3 3 3 ...
##  $ ahi02       : int  3 3 4 4 4 4 3 3 3 3 ...
##  $ ahi03       : int  2 4 3 4 4 4 2 3 2 3 ...
##  $ ahi04       : int  3 3 4 4 4 4 3 4 4 4 ...
##  $ ahi05       : int  3 3 2 3 3 4 3 2 2 4 ...
##  $ ahi06       : int  2 4 3 3 3 4 3 3 3 4 ...
##  $ ahi07       : int  3 4 4 4 4 4 3 3 4 4 ...
##  $ ahi08       : int  3 3 3 4 3 3 3 4 3 3 ...
##  $ ahi09       : int  3 3 3 4 4 4 4 3 4 4 ...
##  $ ahi10       : int  2 2 3 3 4 4 4 3 3 3 ...
##  $ ahi11       : int  3 2 2 3 3 4 4 3 2 3 ...
##  $ ahi12       : int  3 3 3 4 4 4 4 3 4 4 ...
##  $ ahi13       : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ ahi14       : int  2 3 3 4 4 4 3 3 3 3 ...
##  $ ahi15       : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ ahi16       : int  3 3 3 4 4 4 4 3 4 4 ...
##  $ ahi17       : int  2 2 3 4 4 4 3 4 3 4 ...
##  $ ahi18       : int  2 3 3 4 4 4 4 4 3 3 ...
##  $ ahi19       : int  3 3 3 4 4 4 4 2 4 4 ...
##  $ ahi20       : int  3 3 3 4 3 4 4 4 3 4 ...
##  $ ahi21       : int  2 3 2 4 4 4 3 4 3 4 ...
##  $ ahi22       : int  2 3 2 3 4 4 3 3 3 4 ...
##  $ ahi23       : int  3 4 4 4 4 4 3 3 3 3 ...
##  $ ahi24       : int  2 2 3 4 4 4 3 3 4 3 ...
##  $ cesd01      : int  2 2 1 3 1 1 2 2 1 1 ...
##  $ cesd02      : int  1 1 1 2 1 1 3 1 1 1 ...
##  $ cesd03      : int  1 1 1 1 1 1 2 2 1 1 ...
##  $ cesd04      : int  4 4 1 3 1 1 1 4 4 4 ...
##  $ cesd05      : int  1 1 1 1 1 1 1 2 2 3 ...
##  $ cesd06      : int  2 1 1 1 1 1 2 1 1 1 ...
##  $ cesd07      : int  1 2 1 2 1 1 2 1 1 1 ...
##  $ cesd08      : int  3 4 1 1 1 4 4 2 4 4 ...
##  $ cesd09      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cesd10      : int  1 2 1 1 1 1 1 1 1 1 ...
##  $ cesd11      : int  3 2 2 1 3 2 2 2 2 2 ...
##  $ cesd12      : int  2 4 4 3 4 1 3 4 4 4 ...
##  $ cesd13      : int  2 1 1 1 3 1 3 3 1 2 ...
##  $ cesd14      : int  3 2 1 1 1 1 1 3 2 2 ...
##  $ cesd15      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cesd16      : int  2 3 4 3 1 3 3 4 4 4 ...
##  $ cesd17      : int  1 1 1 1 1 1 1 2 1 1 ...
##  $ cesd18      : int  1 1 1 1 1 1 2 2 1 1 ...
##  $ cesd19      : int  2 1 1 1 1 1 1 1 1 1 ...
##  $ cesd20      : int  2 1 1 1 1 1 1 1 1 1 ...
##  $ ahiTotal    : int  63 73 73 89 89 93 80 77 77 85 ...
##  $ cesdTotal   : int  14 6 7 10 13 8 15 12 3 5 ...
```

The second dataset (`participant-info.csv`) contains demographic information about the each of the 295 participants. The data are suitable for various time-series analyses and between-group comparisons. It contains the following variables:

- `id`: Participant's ID.
- `intervention`: 3 positive psychology interventions (PPIs), plus 1 control condition: 1 = Using signature strengths, 2 = Three good things, 3 = Gratitude visit, 4 = Recording early memories (control condition).
- `sex`: 1 for female, 2 for male
- `age`: Participant's age (in years).
- `educ`: Level of education: 1 = Less than Year 12, 2 = Year 12, 3 = Vocational training, 4 = Bachelor's degree, 5 = Postgraduate degree.
- `income`: 1 = below average, 2 = average, 3 = above average.

Let's load our data sets:

```
dat_part <- read.csv("participant-info.csv")
str(dat_part)
```

```
## 'data.frame':    295 obs. of  6 variables:
##  $ id          : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ intervention: int  4 1 4 3 2 1 3 2 1 2 ...
##  $ sex         : int  2 1 1 1 2 1 1 1 1 1 ...
##  $ age         : int  35 59 51 50 58 31 44 57 36 45 ...
##  $ educ        : int  5 1 4 5 5 5 5 4 4 4 ...
##  $ income      : int  3 1 3 2 2 1 2 2 3 3 ...
```

First of all, we must join the two data sets together. We can use the `tidytable` package (which we already load through the `tidyverse` package). The following options are available to join two data sets:

- Considering one data set we add the other one to the left (i.e., we consider the observations in the first data set): `left_join`
- Considering one data set we add the other one to the right (i.e., we consider the observations in the second data set): `right_join`
- Considering both data sets we take the intersection (i.e., observations that are in both data sets in the same time): `inner_join`
- Considering both data sets we take the union (i.e., observations from all data sets): `full_join`

```
dat_full <- tidytable::inner_join(dat, dat_part, by = c("id", "intervention"))
```

After joining two data sets, it is a good practice to check the dimensions of your new data set

```
dim(dat)
```

```
## [1] 992  50
```

```
dim(dat_part)
```

```
## [1] 295   6
```

```
dim(dat_full)
```

```
## [1] 992  54
```

Now, let's see the structure of the data set:

```
str(dat_full)
```

```
## Classes 'tidytable', 'tbl', 'data.table' and 'data.frame':   992 obs. of  54 variables:
```

```
##  $ id          : int  1 1 2 2 2 2 2 2 3 3 ...
##  $ occasion    : int  0 1 0 1 2 3 4 5 0 2 ...
##  $ elapsed.days: num  0 11.77 0 8.02 14.3 ...
##  $ intervention: int  4 4 1 1 1 1 1 1 4 4 ...
##  $ ahi01       : int  2 3 3 3 3 3 3 3 3 3 ...
##  $ ahi02       : int  3 3 4 4 4 4 3 3 3 3 ...
##  $ ahi03       : int  2 4 3 4 4 4 2 3 2 3 ...
##  $ ahi04       : int  3 3 4 4 4 4 3 4 4 4 ...
##  $ ahi05       : int  3 3 2 3 3 4 3 2 2 4 ...
##  $ ahi06       : int  2 4 3 3 3 4 3 3 3 4 ...
##  $ ahi07       : int  3 4 4 4 4 4 3 3 4 4 ...
##  $ ahi08       : int  3 3 3 4 3 3 3 4 3 3 ...
##  $ ahi09       : int  3 3 3 4 4 4 4 3 4 4 ...
##  $ ahi10       : int  2 2 3 3 4 4 4 3 3 3 ...
##  $ ahi11       : int  3 2 2 3 3 4 4 3 2 3 ...
##  $ ahi12       : int  3 3 3 4 4 4 4 3 4 4 ...
##  $ ahi13       : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ ahi14       : int  2 3 3 4 4 4 3 3 3 3 ...
##  $ ahi15       : int  3 3 3 3 3 3 3 3 3 3 ...
##  $ ahi16       : int  3 3 3 4 4 4 4 3 4 4 ...
##  $ ahi17       : int  2 2 3 4 4 4 3 4 3 4 ...
##  $ ahi18       : int  2 3 3 4 4 4 4 4 3 3 ...
##  $ ahi19       : int  3 3 3 4 4 4 4 2 4 4 ...
##  $ ahi20       : int  3 3 3 4 3 4 4 4 3 4 ...
##  $ ahi21       : int  2 3 2 4 4 4 3 4 3 4 ...
##  $ ahi22       : int  2 3 2 3 4 4 3 3 3 4 ...
##  $ ahi23       : int  3 4 4 4 4 4 3 3 3 3 ...
##  $ ahi24       : int  2 2 3 4 4 4 3 3 4 3 ...
##  $ cesd01      : int  2 2 1 3 1 1 2 2 1 1 ...
##  $ cesd02      : int  1 1 1 2 1 1 3 1 1 1 ...
##  $ cesd03      : int  1 1 1 1 1 1 2 2 1 1 ...
##  $ cesd04      : int  4 4 1 3 1 1 1 4 4 4 ...
##  $ cesd05      : int  1 1 1 1 1 1 1 2 2 3 ...
##  $ cesd06      : int  2 1 1 1 1 1 2 1 1 1 ...
##  $ cesd07      : int  1 2 1 2 1 1 2 1 1 1 ...
##  $ cesd08      : int  3 4 1 1 1 4 4 2 4 4 ...
##  $ cesd09      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cesd10      : int  1 2 1 1 1 1 1 1 1 1 ...
##  $ cesd11      : int  3 2 2 1 3 2 2 2 2 2 ...
##  $ cesd12      : int  2 4 4 3 4 1 3 4 4 4 ...
##  $ cesd13      : int  2 1 1 1 3 1 3 3 1 2 ...
##  $ cesd14      : int  3 2 1 1 1 1 1 3 2 2 ...
##  $ cesd15      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ cesd16      : int  2 3 4 3 1 3 3 4 4 4 ...
##  $ cesd17      : int  1 1 1 1 1 1 1 2 1 1 ...
##  $ cesd18      : int  1 1 1 1 1 1 2 2 1 1 ...
##  $ cesd19      : int  2 1 1 1 1 1 1 1 1 1 ...
##  $ cesd20      : int  2 1 1 1 1 1 1 1 1 1 ...
##  $ ahiTotal    : int  63 73 73 89 89 93 80 77 77 85 ...
##  $ cesdTotal   : int  14 6 7 10 13 8 15 12 3 5 ...
##  $ sex         : int  2 2 1 1 1 1 1 1 1 1 ...
##  $ age         : int  35 35 59 59 59 59 59 59 51 51 ...
##  $ educ        : int  5 5 1 1 1 1 1 1 4 4 ...
##  $ income      : int  3 3 1 1 1 1 1 1 3 3 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

We must make some preprocessing:

- Transform some variable as factor ones: occasion, intervention, sex, educ, and income
- Check for NAs and/or outliers
- Remove the all the ahi and cesd variables except for the total variables.

```
ahi_var <- colnames(dat_full)[grepl("ahi", colnames(dat_full))]
cesd_var <- colnames(dat_full)[grepl("cesd", colnames(dat_full))]

dat_full <- dat_full %>%
  dplyr::select(-c(ahi_var[-25], cesd_var[-21])) %>%
  mutate(occasion = as.factor(occasion),
         intervention = as.factor(intervention),
         sex = as.factor(sex),
         educ = as.factor(educ),
         income = as.factor(income))
sum(is.na(dat_full))
```

```
## [1] 0
```

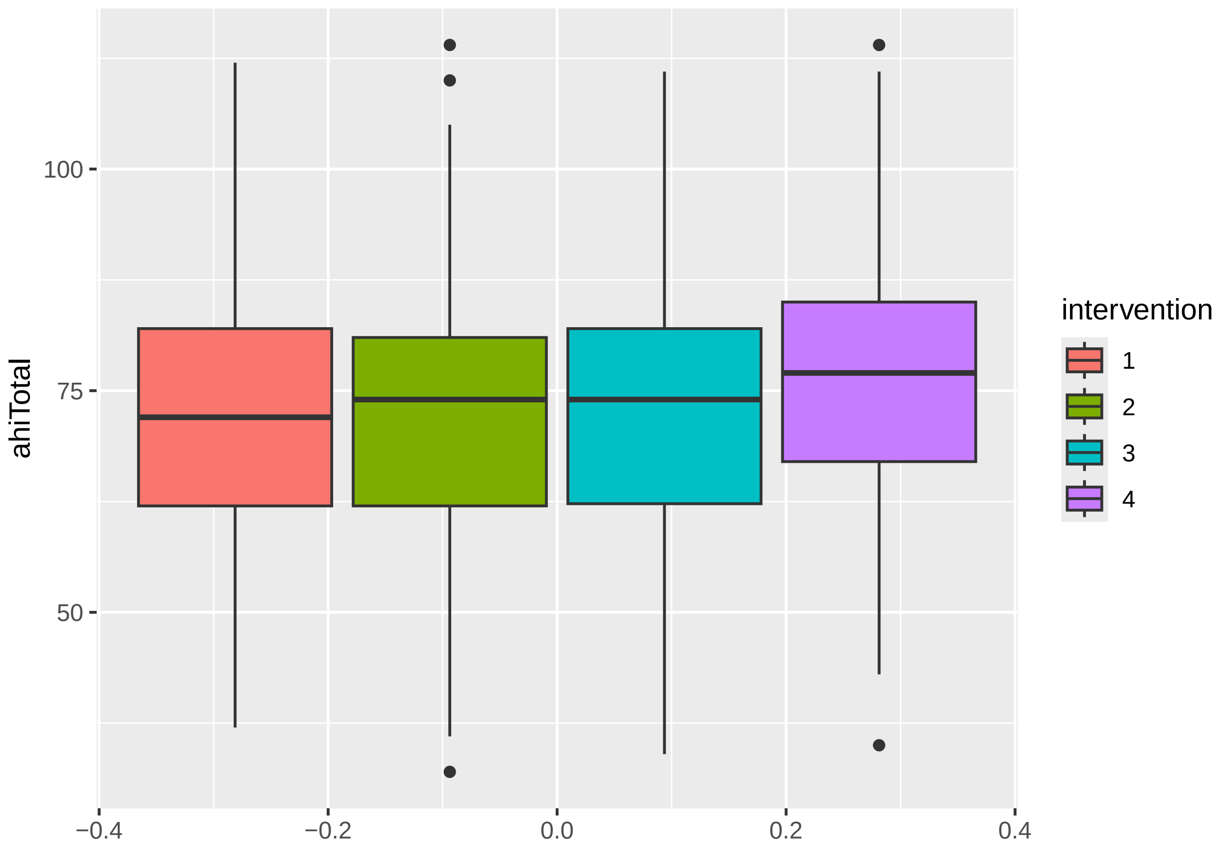
```
summary(dat_full)
```

```
##        id          occasion  elapsed.days    intervention    ahiTotal
##  Min.   :  1.00   0:295    Min.   :  0.00   1:232        Min.   : 32.00
##  1st Qu.: 74.75   1:147    1st Qu.:  0.00   2:289        1st Qu.: 63.00
##  Median :147.00   2:157    Median : 14.79   3:210        Median : 74.00
##  Mean   :147.36   3:139    Mean   : 44.31   4:261        Mean   : 72.79
##  3rd Qu.:218.25   4:134    3rd Qu.: 90.96                3rd Qu.: 83.00
##  Max.   :295.00   5:120    Max.   :223.82                Max.   :114.00
##    cesdTotal     sex          age         educ    income
##  Min.   : 0.00   1:843   Min.   :18.00   1: 47   1:246
##  1st Qu.: 4.00   2:149   1st Qu.:35.00   2: 83   2:447
##  Median :10.00           Median :46.00   3:131   3:299
##  Mean   :13.14           Mean   :45.04   4:325
##  3rd Qu.:19.00           3rd Qu.:54.00   5:406
##  Max.   :55.00           Max.   :83.00
```

There are no missing values.

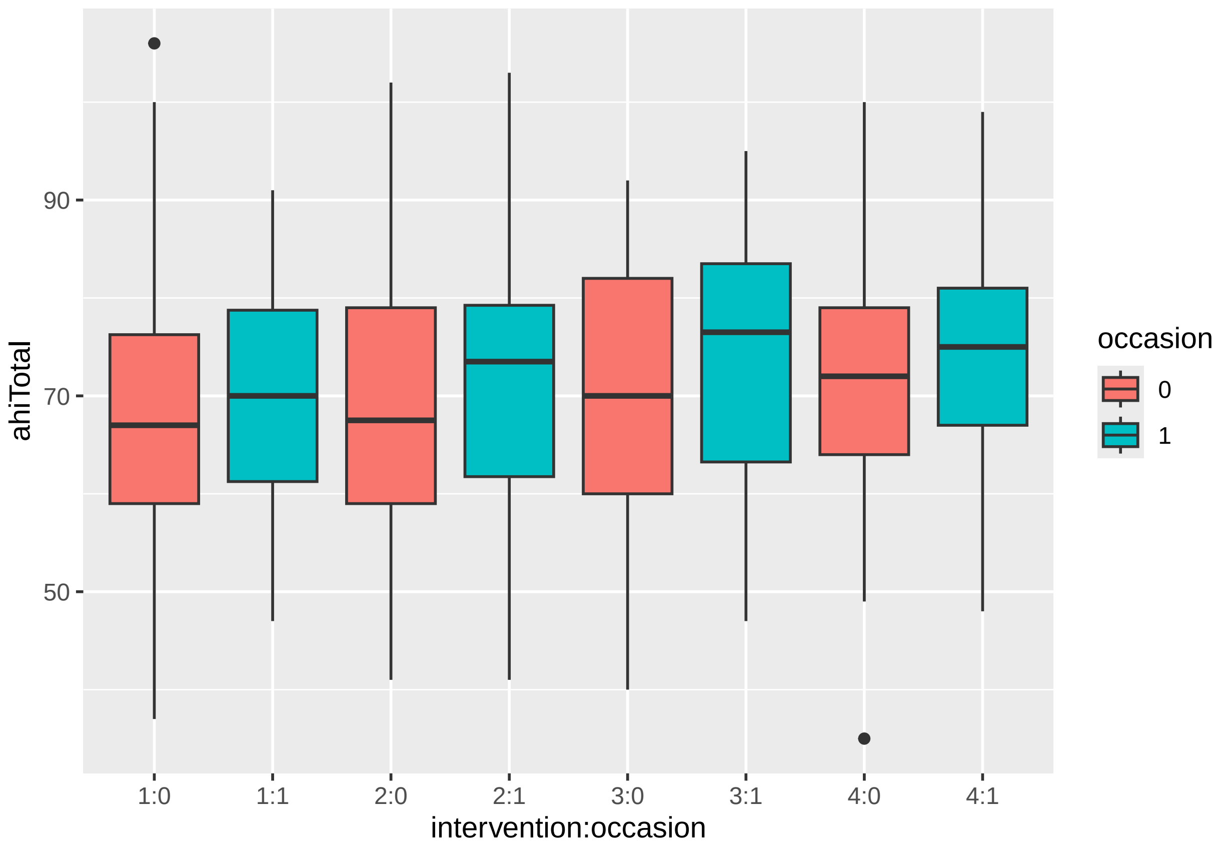
Now, we can create some exploratory plots. Let's see the distribution of the total AHI score divided by type of intervention:

```
ggplot(dat_full) +
  geom_boxplot(aes(y = ahiTotal, fill = intervention))
```



However, we want to see if the total AHI score increases after the intervention. For simplicity, let's consider the first and second occasions (i.e., occasion equals 0 and 1):

```
dat_full %>%
  filter(occasion %in% c(0,1)) %>%
  ggplot() +
  geom_boxplot(aes(y = ahiTotal, x = intervention:occasion, fill = occasion))
```

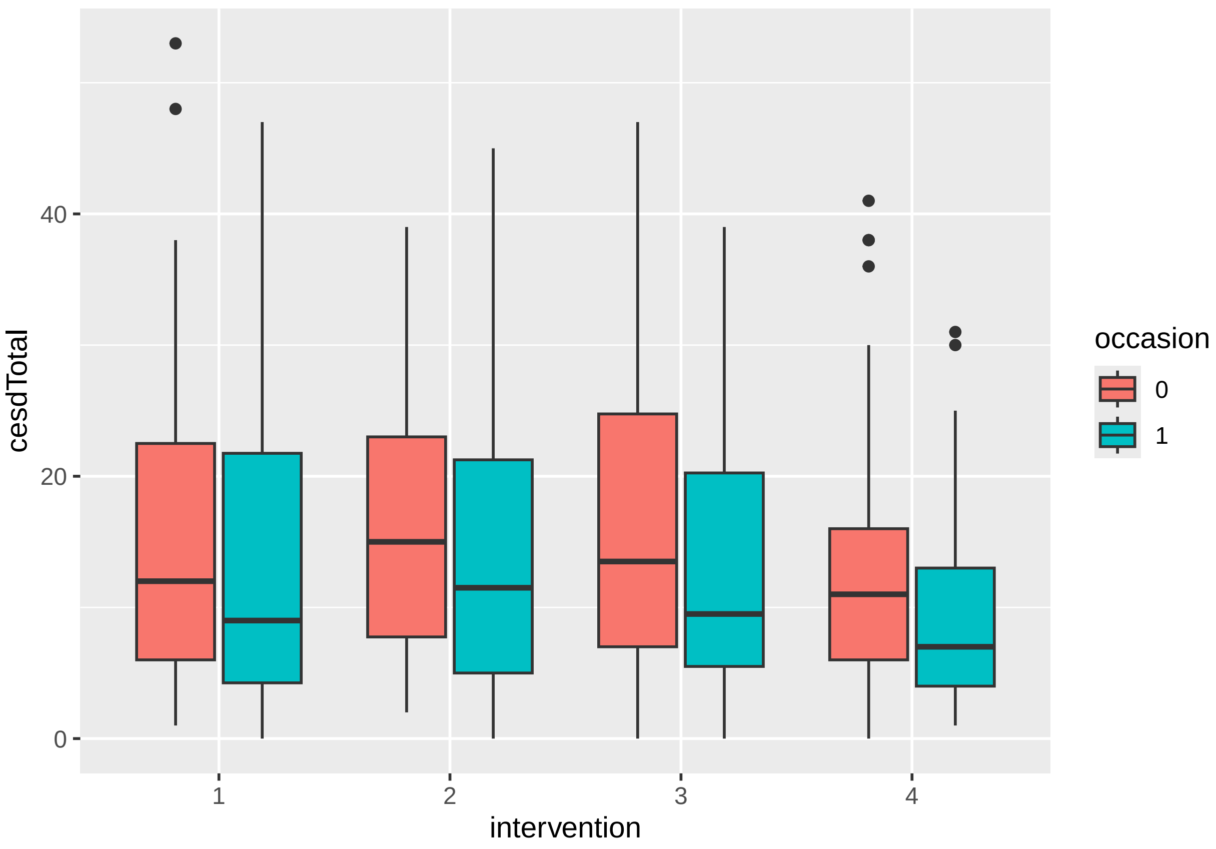


```
#geom_boxplot(aes(y = ahiTotal, x = intervention:occasion, fill = occasion))
```

We can note an increment of the total AHI score for all the type of intervetion.

Let's analyze the total CESD one:

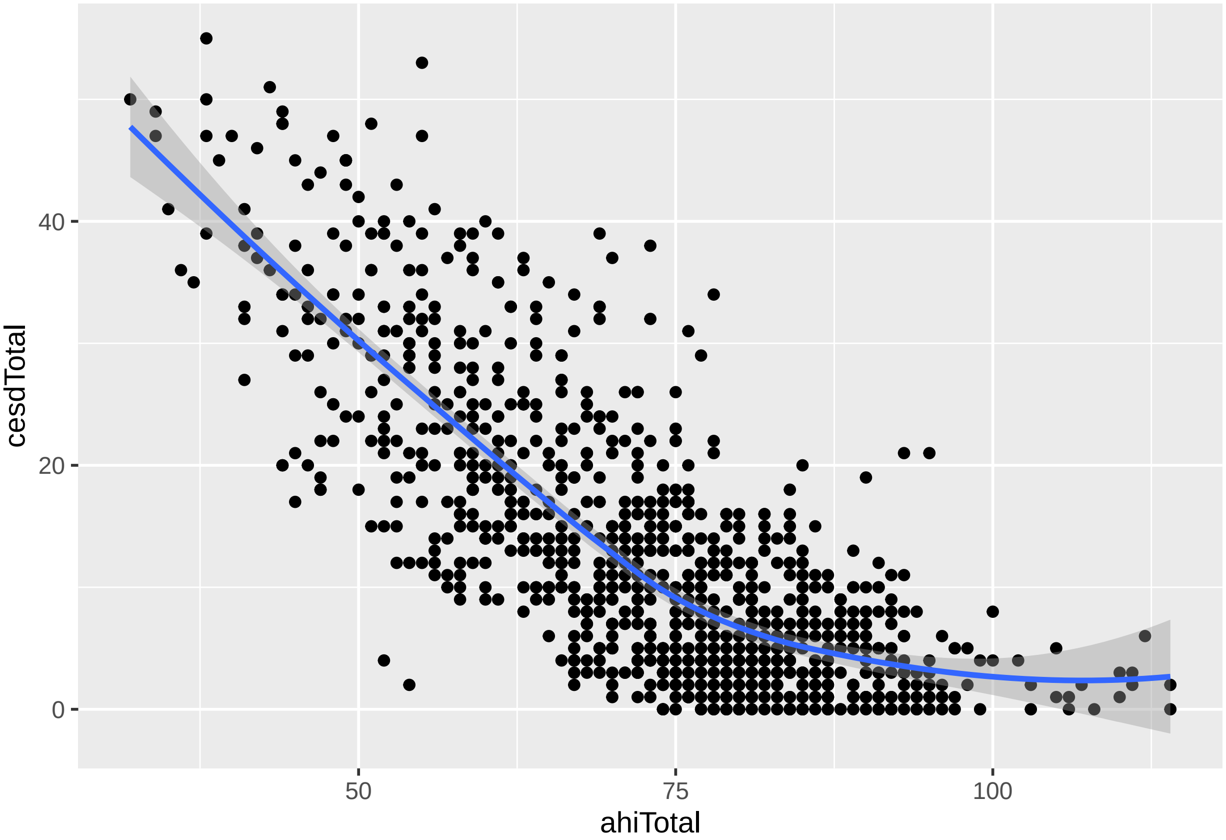
```
dat_full %>%
  filter(occasion %in% c(0,1)) %>%
  ggplot() +
  geom_boxplot(aes(y = cesdTotal, x = intervention, fill = occasion))
```



Here, we can see a reduction of the total CESD score for all type of psychological intervention.

Let's now explore the relationship between the total AHI score and total CESD score

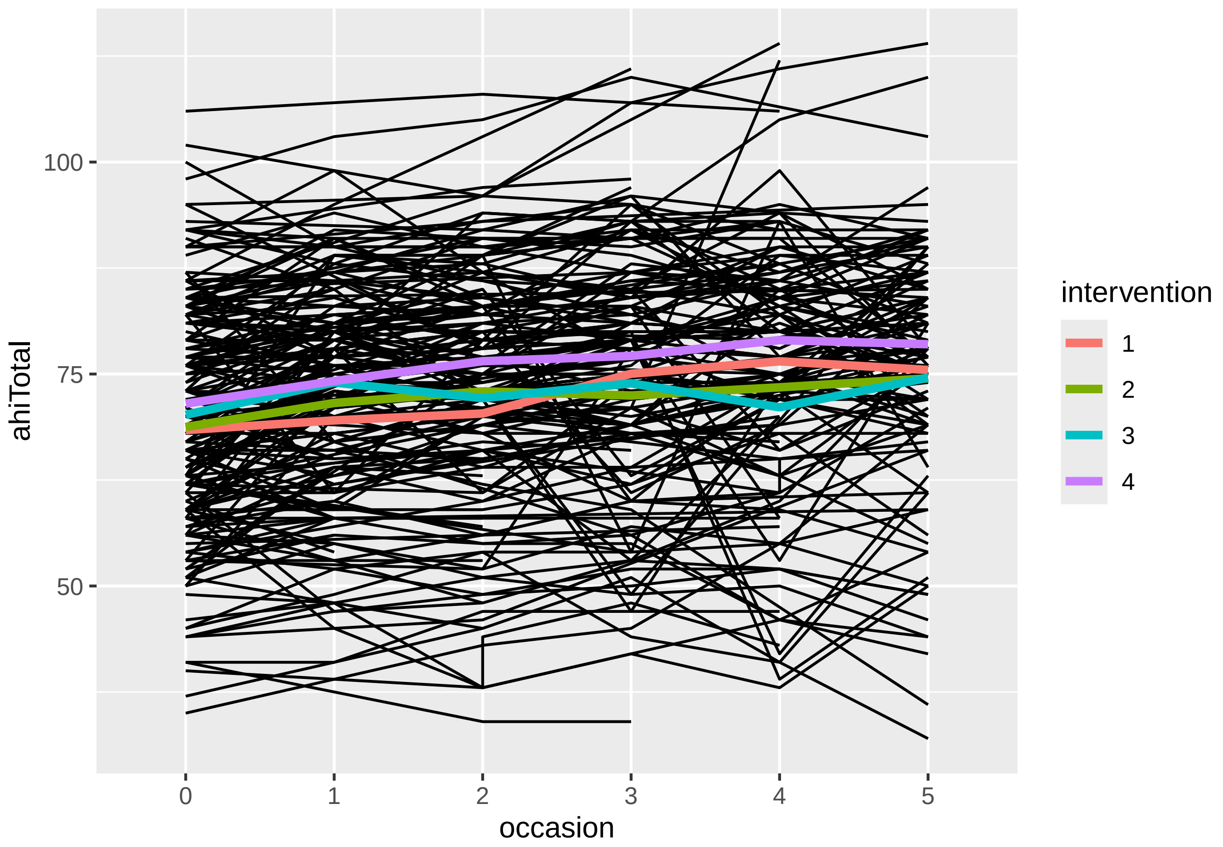
```r
ggplot(dat_full) +
  geom_point(aes(x = ahiTotal, y = cesdTotal)) +
  geom_smooth(aes(x = ahiTotal, y = cesdTotal), method = 'loess')
```



So, high values of AHI correspond to low value of CESD in general.

Another interesting point is to see the total AHI score for each timepoint and each participants and the mean for each intervention:

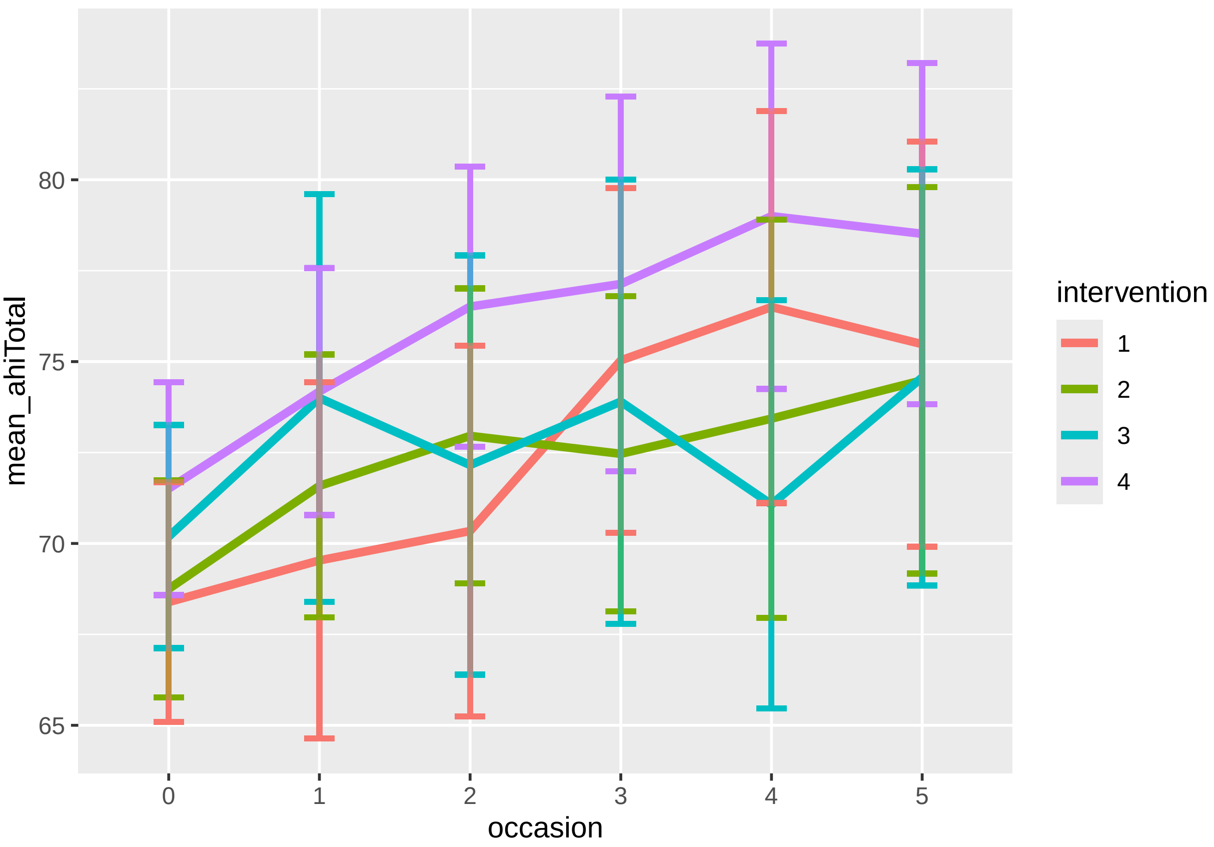
```
dat_full %>%
  group_by(intervention, occasion) %>%
  mutate(mean_ahiTotal = mean(ahiTotal)) %>%
  ggplot() +
  geom_line(aes(x = occasion, y = ahiTotal, group = id)) +
    geom_line(aes(x=occasion,
           y=mean_ahiTotal,
           group=intervention,
           colour=intervention), linewidth=1.5)
```



or considering directly the global mean for each occasion and intervention with corresponding 0.95 confidence intervals:

```
dat_full %>%
  group_by(intervention, occasion) %>%
  mutate(mean_ahiTotal = mean(ahiTotal),
         sd_ahiTotal=sd(ahiTotal),
         n_ahiTotal=length(ahiTotal),
         upper=mean_ahiTotal+2*sd_ahiTotal/sqrt(n_ahiTotal),
         lower=mean_ahiTotal-2*sd_ahiTotal/sqrt(n_ahiTotal)) %>%
  ggplot() +
      geom_line(aes(x=occasion,
             y=mean_ahiTotal,
             group=intervention,
             colour=intervention), size=1.5) +
    geom_errorbar(aes(x=occasion, ymin=upper, ymax=lower,
                      color=intervention),
                  width=0.2, linewidth=1,alpha=.5)
```

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use `linewidth` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```



Let's save our preprocessed data set as RData file for the next lesson

```
save(dat_full, file = "dat_full.RData")
```

Seligman, Steen, M. E. P. 2005. "Positive Psychology Progress: Empirical Validation of Interventions." *American Psychologist* 60: 410–21. https://doi.org/https://doi.org/10.1037/0003-066X.60.5.410.

Woodworth, Rosalind J., Angela O'Brien-Malone, Mark R. Diamond, and Benjamin Schüz. 2017. "Web-Based Positive Psychology Interventions: A Reexamination of Effectiveness." *Journal of Clinical Psychology* 73 (3): 218–32. https://doi.org/https://doi.org/10.1002/jclp.22328.