Analisis of access of education and training for persons with disability in Europe

Bianca Isabel

7/11/2021

Packages used:

- dplyr
- ggplot

Data preparation and introduction

-Original data from: https://ec.europa.eu/eurostat/data/database

-Modified CSV from: https://www.kaggle.com/gpreda/access-to-education-of-disabled-people-in-europe

The data presents the results of the evaluation for the accessibility to education and training for persons with disabilities in EU, reported in Eurostat, the evaluation is done with the next parameters:

1. Units: all in thousands, not printed for cleanness
2. ISCE97: International Standard Classification of Education 1997

- ED0-2: Pre-primary to low secondary education
- ED3-4: High secondary to Post secondary
- ED5-6: First and second part of tertiary education
- NRP: Not reported

3. HLTH_PB: European disability level classification

- PB1040 - Difficulty in basic activities
- PB1041 - No difficulty in basic activities
- PB1070 - Limitation in work caused by a health condition or difficulty in a basic activity
- PB1071 - No limitation in work caused by a health condition or difficulty in basic activities
- TOTAL - Sum of all the disability levels classification
- NRP - Not reported

4. Sex

- M - Males
- F - Female
- T - Sum of M and F

5. Age group

- 15-24
- 25-34
- 35-44
- 45-54
- 55-64
- Total - Sum of all age group

6. Time: year of evaluation, all in 2011

7. Geo: two letter code of country
8. Value: numerical value of examination done for accessibility of education and training for persons with disability in Europe

Loading and cleaning of the data:

```
# Data loading
eu_ed=read.csv('education_disbled_eu.csv')

# Cleaning of data
eu_ed_nt = mutate_if( #make strings factors
  subset(eu_ed, select = -c(unit,time), #take out units and year as is the same in all
         eu_ed$sex!="T" & eu_ed$age!="TOTAL" & eu_ed$hlth_pb!="TOTAL" & eu_ed$isced97!="TOTAL"),#with o
  is.character, as.factor)

eu_ed_ms = eu_ed_nt[c(which(complete.cases(eu_ed_nt)==FALSE)),] #table of missing reported examination
eu_ed_cm = na.omit(eu_ed_nt) #table of does with examination values
```

Let us first analyze the general data set to obtain an idea of the whole:

```
## data summary
summary.data.frame(eu_ed_nt)
```

```
##   isced97       hlth_pb     sex          age           geo
##  ED0-2:1240   PB1040:1240   F:2480   Y15-24:992   AT     : 160
##  ED3_4:1240   PB1041:1240   M:2480   Y25-34:992   BE     : 160
##  ED5_6:1240   PB1070:1240            Y35-44:992   BG     : 160
##  NRP  :1240   PB1071:1240            Y45-54:992   CH     : 160
##                                      Y55-64:992   CY     : 160
##                                                   CZ     : 160
##                                                   (Other):4000
##      value
##  Min.   :   0.506000
##  1st Qu.:  12.560000
##  Median :  51.405500
##  Mean   : 209.769553
##  3rd Qu.: 195.248750
##  Max.   :3954.942000
##  NA's   :1494
```

```
summary.data.frame(eu_ed_cm)
```

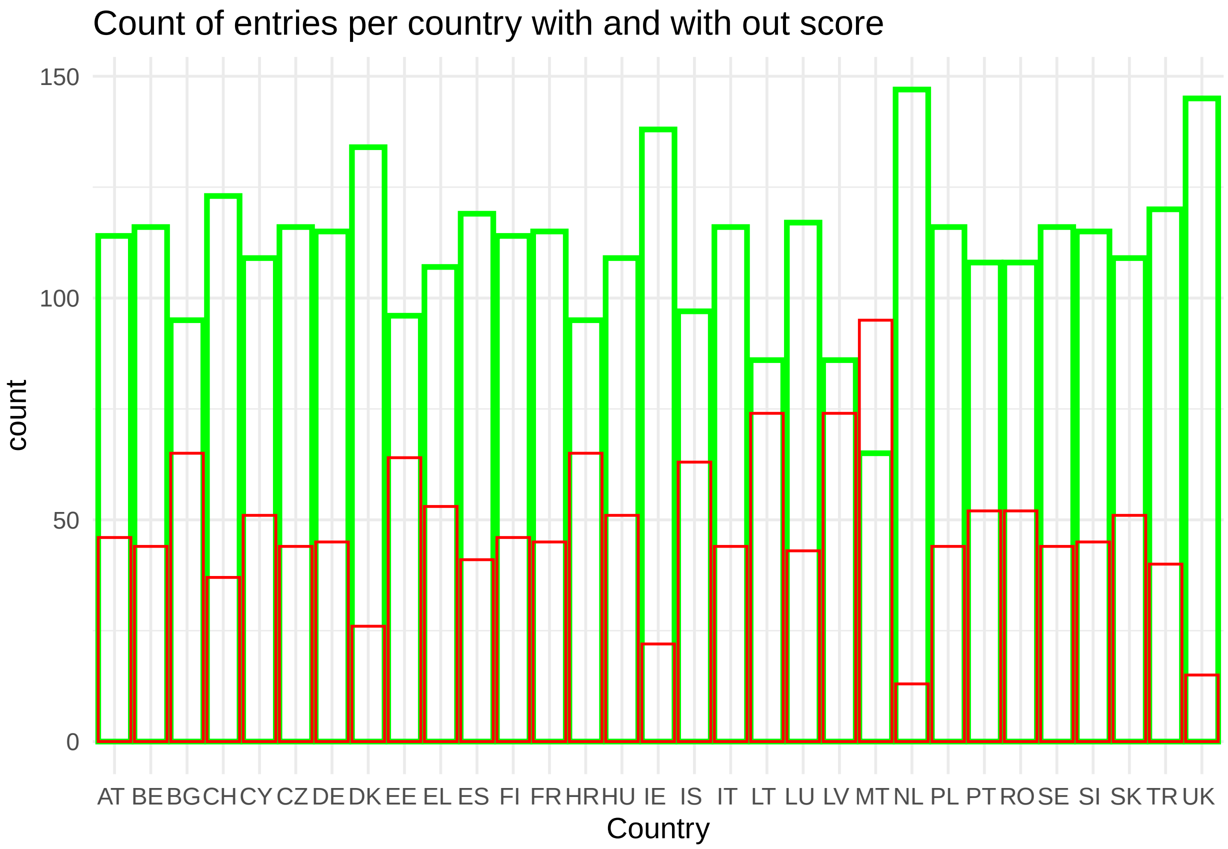
```
##   isced97       hlth_pb    sex          age           geo
##  ED0-2:1171   PB1040:777   F:1743   Y15-24:588   NL     : 147
##  ED3_4:1172   PB1041:968   M:1723   Y25-34:680   UK     : 145
##  ED5_6:1017   PB1070:752            Y35-44:716   IE     : 138
##  NRP  : 106   PB1071:969            Y45-54:730   DK     : 134
##                                     Y55-64:752   CH     : 123
##                                                  TR     : 120
##                                                  (Other):2659
##      value
##  Min.   :   0.506000
##  1st Qu.:  12.560000
##  Median :  51.405500
##  Mean   : 209.769553
##  3rd Qu.: 195.248750
##  Max.   :3954.942000
```

```
##
```

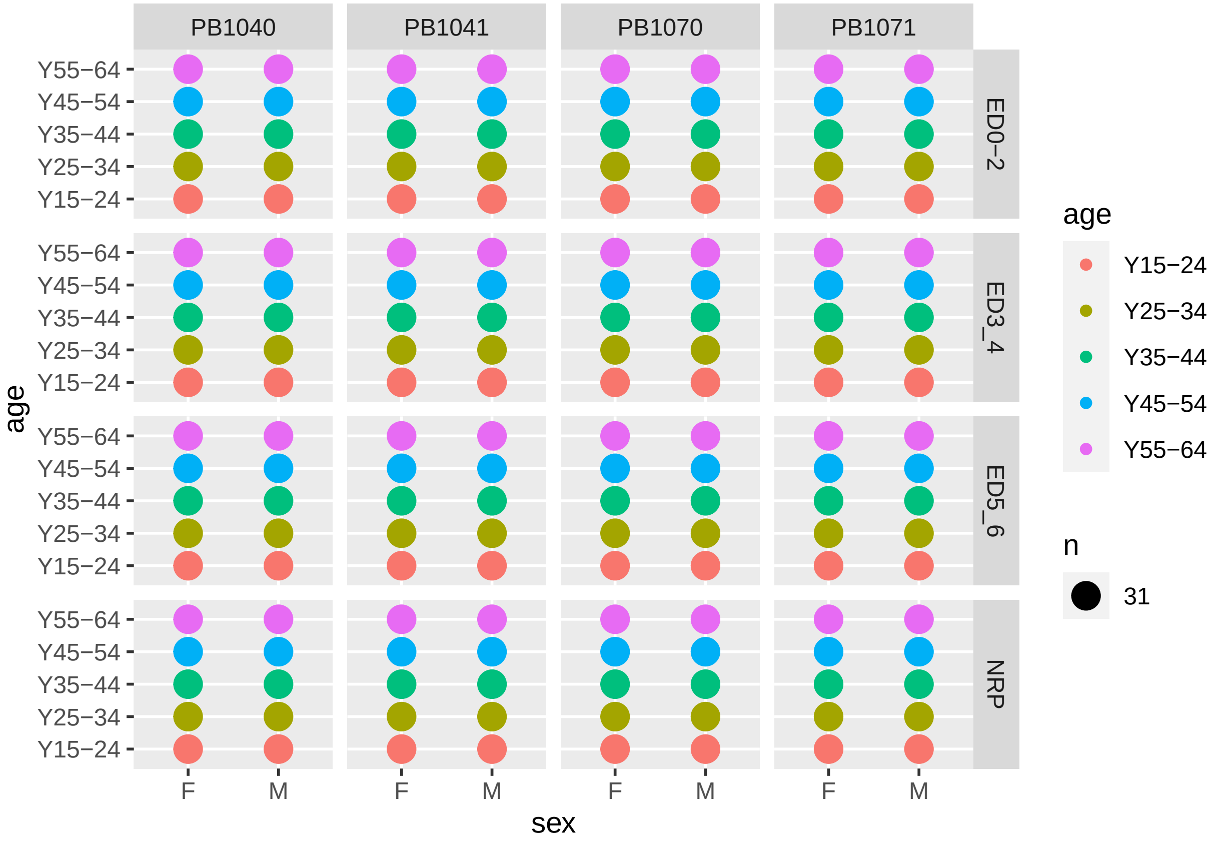
```
summary.data.frame(eu_ed_ms)
```

```
##   isced97       hlth_pb     sex         age             geo             value
##  ED0-2:  69   PB1040:463   F:737   Y15-24:404   MT      :  95   Min.   : NA
##  ED3_4:  68   PB1041:272   M:757   Y25-34:312   LT      :  74   1st Qu.: NA
##  ED5_6: 223   PB1070:488           Y35-44:276   LV      :  74   Median : NA
##  NRP  :1134   PB1071:271           Y45-54:262   BG      :  65   Mean   :NaN
##                                    Y55-64:240   HR      :  65   3rd Qu.: NA
##                                                 EE      :  64   Max.   : NA
##                                                 (Other):1057   NA's   :1494
```

```
## bars of does with scores (green) and with out score (red)
ggplot(eu_ed_cm, aes(x=geo))+
  geom_bar(colour="green", fill=NA, position="stack", size=1)+
  geom_bar(data = eu_ed_ms, colour="red", fill=NA, position = "stack")+
  labs(x="Country")+
  ggtitle("Count of entries per country with and with out score")+
  theme(legend.position = "none")+
  theme_minimal()
```



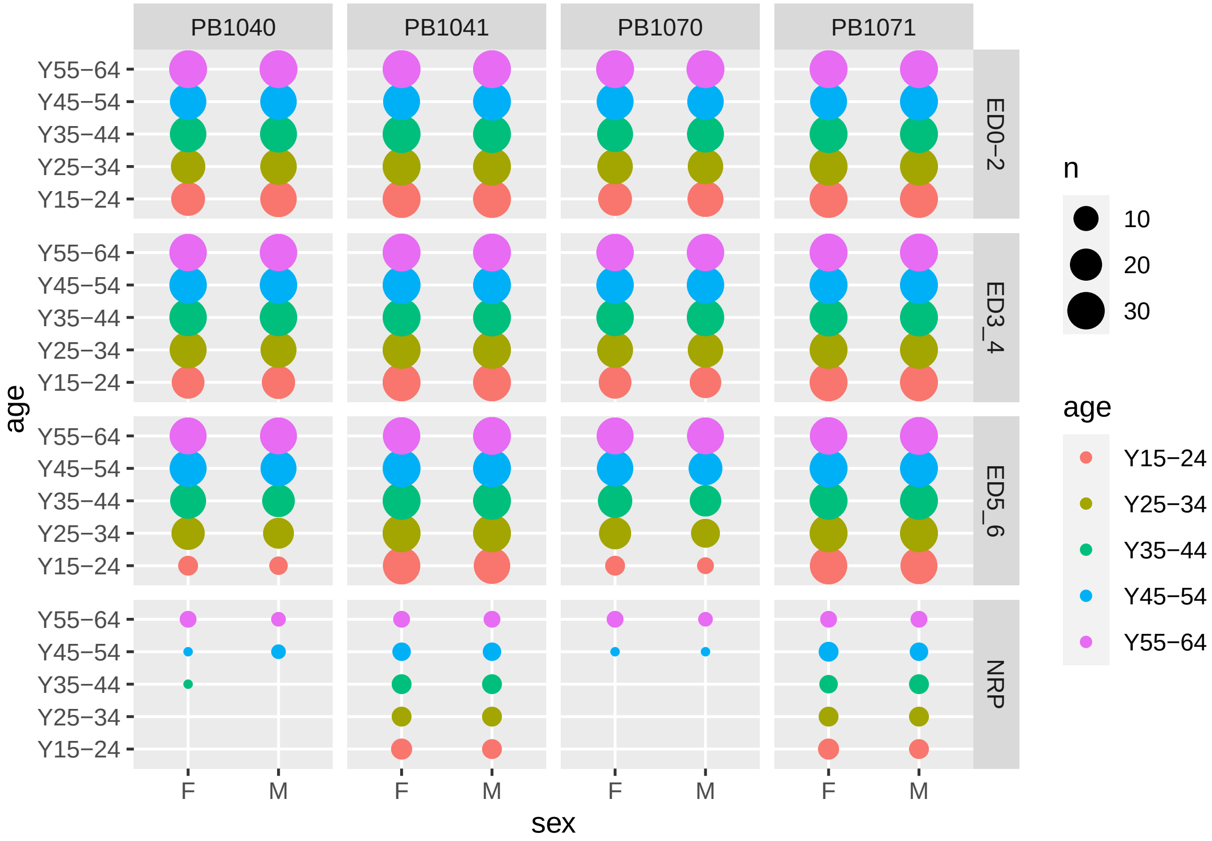
```
# comparing all of the data of education level with reported scholarship sex, and age group
ggplot(eu_ed_nt, aes(sex,age, colour=age))+
  geom_count() +
  facet_grid(rows = vars(isced97), cols = vars(hlth_pb))
```



```
  theme(title="Analysis of the whole data set, with respect with respect to demographics")
```

```
## List of 1
##  $ title: chr "Analysis of the whole data set, with respect with respect to demographics"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

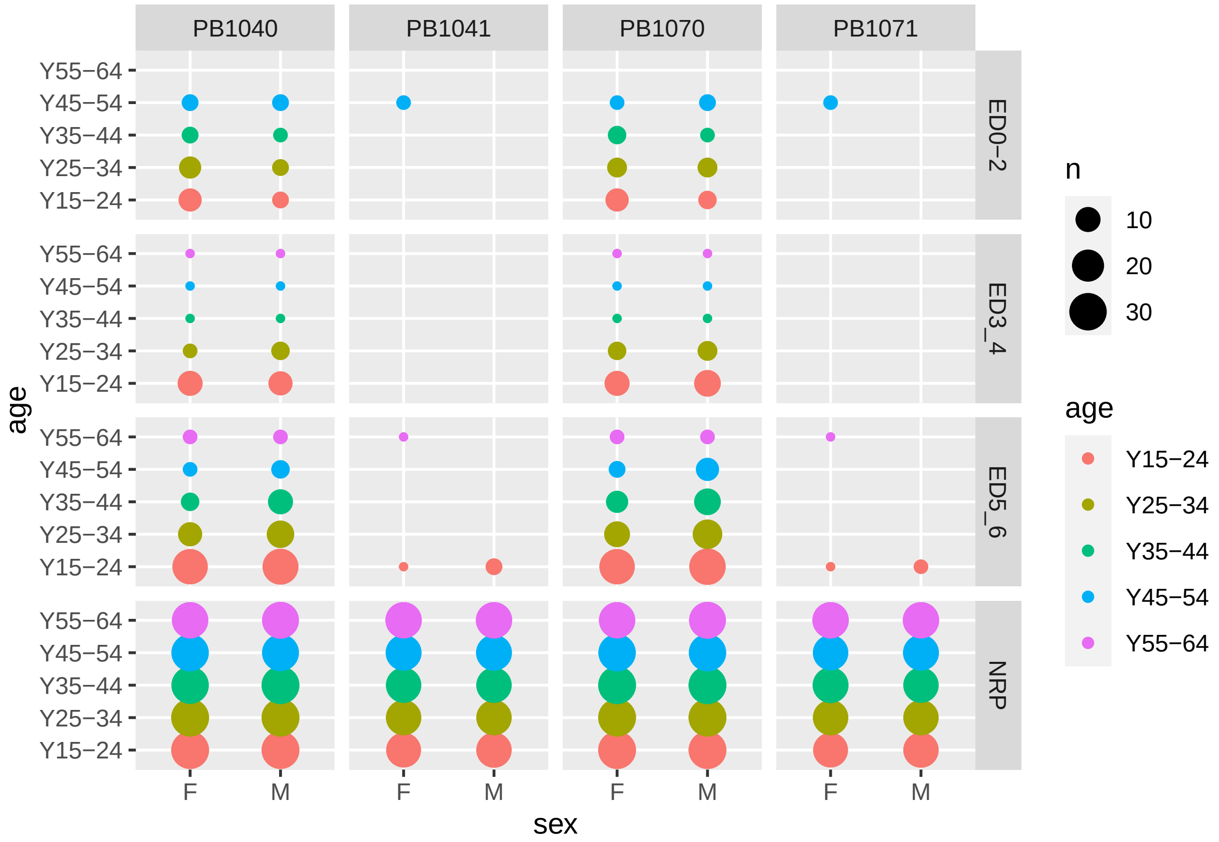
```
# comparing cm data of education level with reported scholarship sex, and age group
ggplot(eu_ed_cm, aes(sex,age, colour=age))+
  geom_count() +
  facet_grid(rows = vars(isced97), cols = vars(hlth_pb))
```



```
  theme(title = "Analysis of does with reported scores with respect to demographics")
```

```
## List of 1
##  $ title: chr "Analysis of does with reported scores with respect to demographics"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

```
# comparing ms data of education level with reported scholarship sex, and age group
ggplot(eu_ed_ms, aes(sex,age, colour=age))+
  geom_count() +
  facet_grid(rows = vars(isced97), cols = vars(hlth_pb))
```



```
  theme(title ="Analysis of does with missing scores with respect to demographics")
```

```
## List of 1
##  $ title: chr "Analysis of does with missing scores with respect to demographics"
##  - attr(*, "class")= chr [1:2] "theme" "gg"
##  - attr(*, "complete")= logi FALSE
##  - attr(*, "validate")= logi TRUE
```

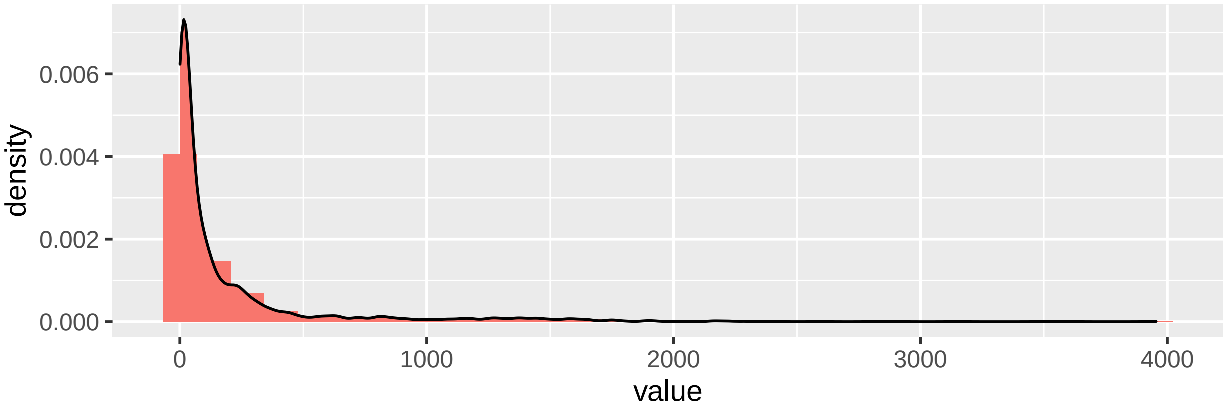
As we can see the only country with more missing scores is MT. The ones with more scores are UK and NL, with NL being the one with less missing scores. In age must scores comes from ages 55 to 64.

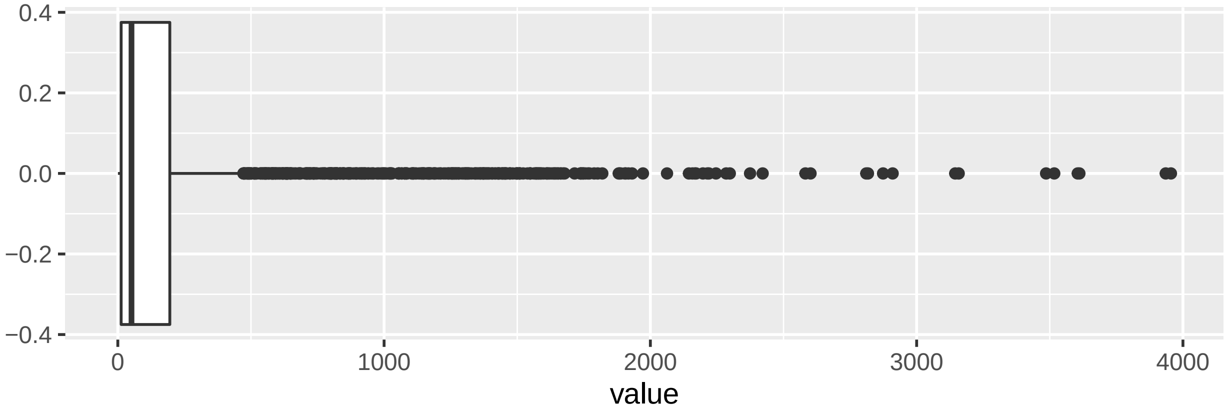
Anlyzing distribiutions

First let us see the general distribution of the scores:

```
## distribution with histogram
p1=ggplot(eu_ed_cm, aes(value, fill="#3590e0"))+
  geom_histogram(aes(y=..density..))+
  geom_density()+
  theme(legend.position = "none")

## box plot
p2=ggplot(eu_ed_cm, aes(value))+
  geom_boxplot()

multiplot(p1,p2, layout = matrix(c(1,2)) )
```

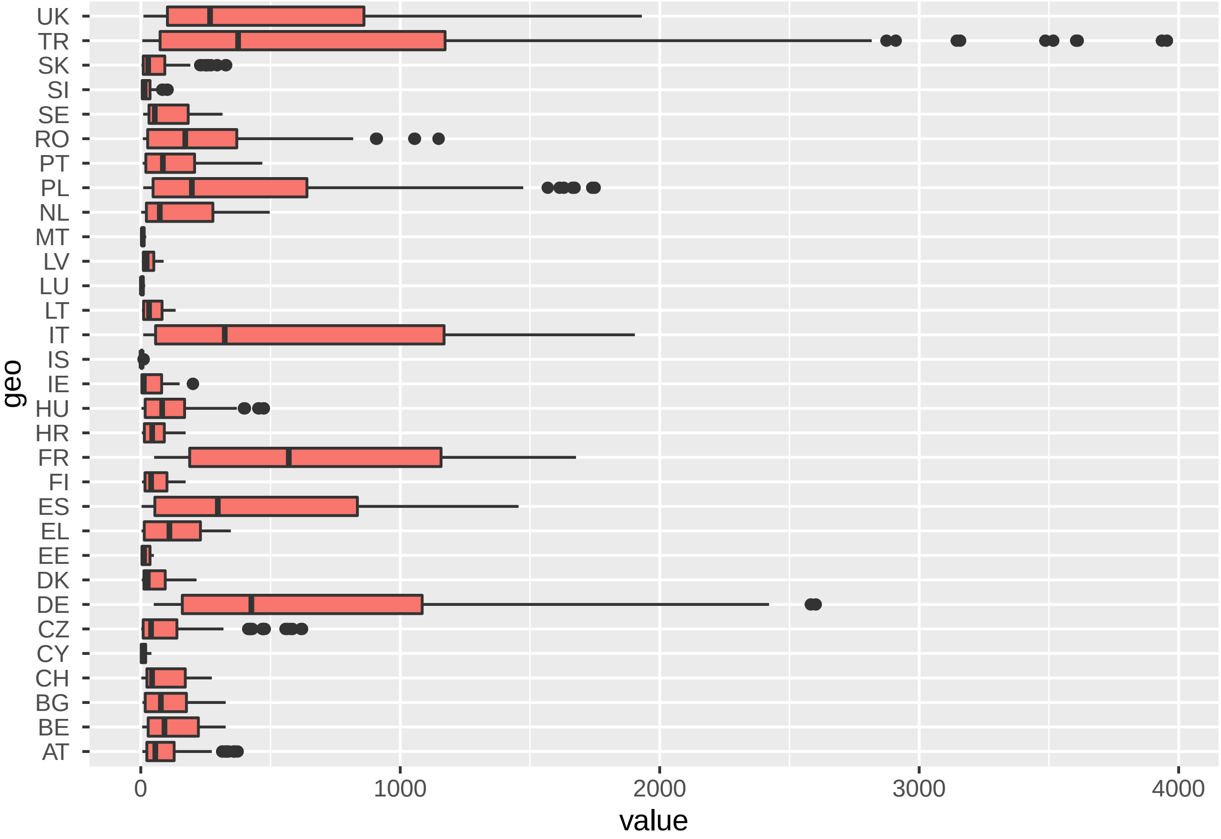
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



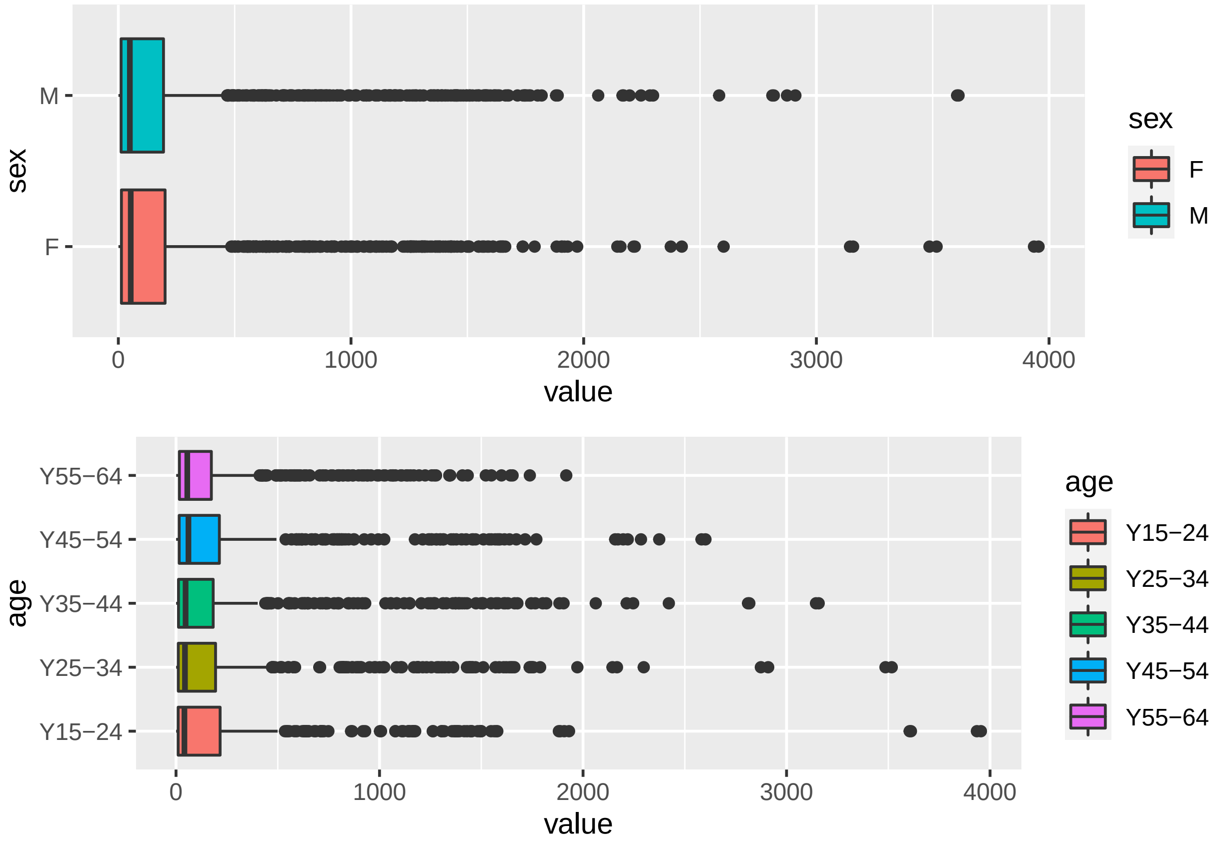


As we can see the values go to density close to 0 after 500, and the box plot shows a lot of outliers, lets see if they still there if broken down by category

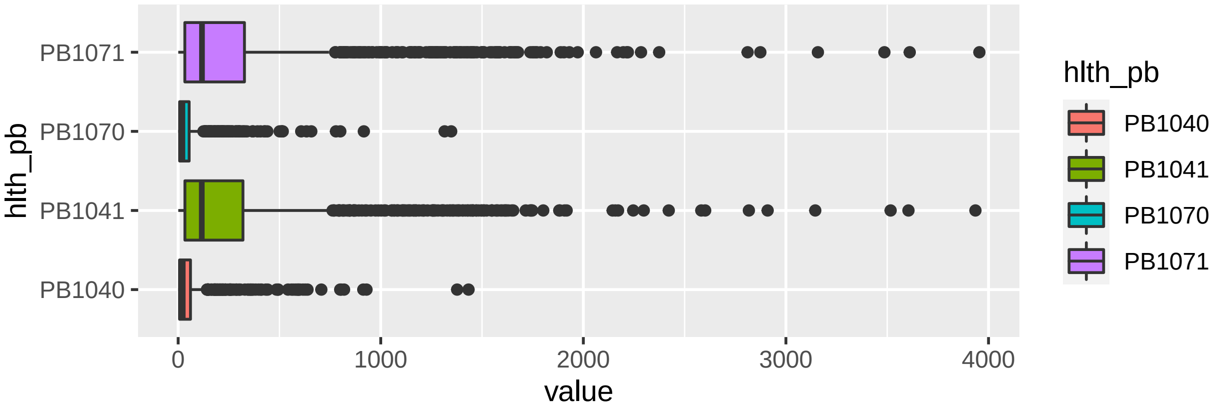
```
# Using eu_ed_cm to remove warning about removed rows

# Value with respect to country
p1=ggplot(eu_ed_cm, aes(value, geo, fill="red"))+
  geom_boxplot()+
  theme(legend.position = "")

# Value with respect to sex
p2=ggplot(eu_ed_cm, aes(value, sex, fill=sex))+
  geom_boxplot()

# Value with respect to age group
p3=ggplot(eu_ed_cm, aes(value, age, fill=age))+
  geom_boxplot()

# Value with respect to disability classification
p4=ggplot(eu_ed_cm, aes(value, hlth_pb, fill=hlth_pb))+
  geom_boxplot()

# Value with respect to education level
p5=ggplot(eu_ed_cm, aes(value, isced97, fill=isced97))+
  geom_boxplot()

## desing
p1
```

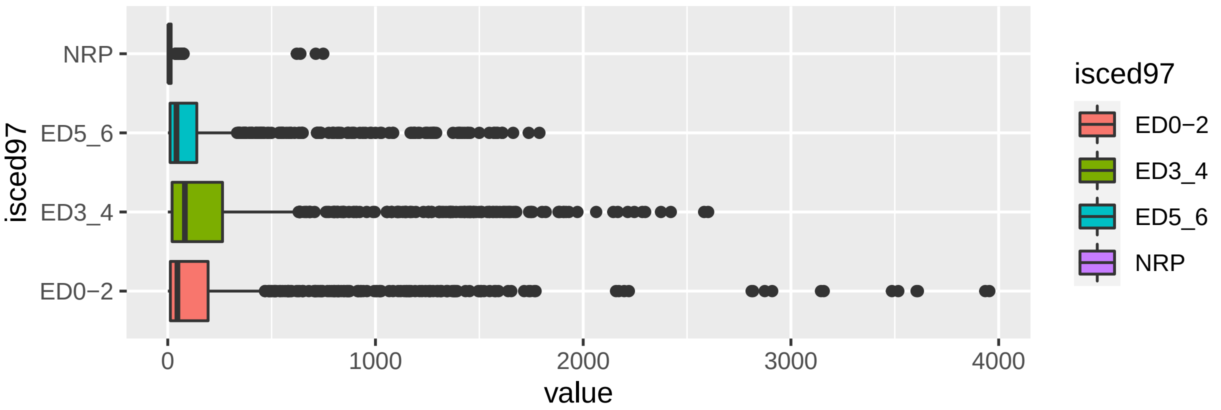


```
multiplot(p2,p3)
```



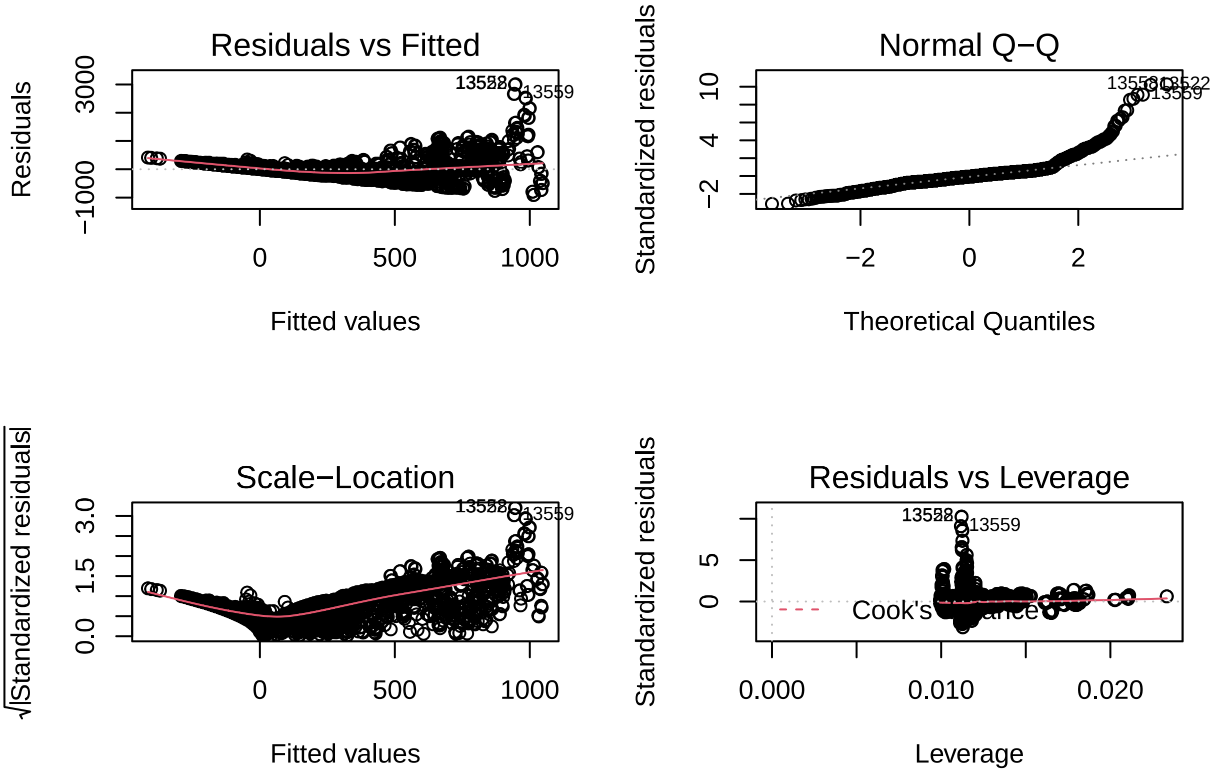
```
multiplot(p4,p5)
```





The outlines continue except in all the cases except in the country where they are less. We can analyze a linear model to see how our data behaves and to see if there are any outliers that we can take out using cooks distance as a parameter so to make a decision to go foward with the analyzis.

```r
model = lm(value ~ ., data = eu_ed_nt)

par(mfrow = c(2,2))
plot(model)
```



```
ck=cooks.distance(model)
infl=(ck>4*mean(ck, na.rm = T))
eu_ed_nt_no=eu_ed_nt[-infl,]
```