# There are decisions to be made about data

29/03/2016: Some important decsions best made before producing a report which we show to too many people.

## Countries to include.

The big dataset is a little problematic. Some people have chosen countries wehre survey was not done. This really is due to the data entry validation problems in the small datasets. What could their reasons have been - choosing an erroneous country from the list?But possibly because they were coming from that country and were now in the other country where the survey was done.

The number of such cases is small and can be investigated:

Find those where suspiciously low answer counts: `sort(table(droplevels(data_Europe$q_0000)),T)` or `as.data.frame(sort(table(droplevels(data_Europe$q_0000)),T))`

Then give list of actual survey countries: `list.dirs(path="../data/", recursive=T,full.names=F)`

**So suspects to investigate are:**

| Ctry | Count |
|---|---|
| Switzerland | 2 |
| Belarus | 1 |
| Croatia | 1 |
| Serbia | 1 |
| United Kingdom | 1 |

Probably these values can be traced back in the original survey data frames and can be set to correct values either during reading, or used as they are. Well, I am not too sure about that 2nd option.

## Another decision is what to do with NA values.

We probably want to discard data where NA values are present in most cells. At least on the first page of the survey. *Rationale* is - if they have not done mandatory questions on the first page - they could not possibly have done the second page. 1st page questions:

`c("q_0001", "q_0002", "q_0003", "q_0004")`

Theoretically there could be those who abandoned survey on 2nd or 3rd page - but we need to check how many such cases exist. So we would need to check:

- if there aree people who have not entered demographics, but have written something in `q_006`
- if there are people who have done the first page but then abandoned the survey - i.e. who have NAs in 2nd, 3rd or 4th page mandatory questions
- hypothesis is that such number will be very small

There are some columns found in the dataset where there will never be a NA value:

`sort(apply(data_Europe, 2, function(x) length(which(is.na(x)))))`

namely question `q_0005_S*` and `q_0009_S*` options will always be set. So we could choose to discard those in our criteria for completed pages.

**check if thos who have answered in question `q_0006`**

```
nrow(completeFun(data_Europe[!rowSums(is.na(data_Europe[12])),],naS1stpg)) - nrow(data_Europe[!rowSums(
```

well that was a bit silly - I forgot that those who answered this question must have entered something on previous question. however it is still staisfying to obtain answer 0 - which means there is nobody in the data set who has any NAs in first 4 questions who has typed anything in `q_0006`. We can use this methiod for further checks in other pages.