How to use the GSOEP

Michel Lubrano Lara Vivian

February 18, 2015

1 Introduction

First of all, there are two sources of information available to access the GSOEP

1. The SOEP Desktop Companion (DTC) which is available as a pdf file on the DIW web site
   `http://www.diw.de/en/soep`

2. The following web site gives an access to variable and concept definition, year availability

   - `https://data.soep.de/search/concepts`
   - `https://data.soep.de/search/variables`

   This site plays an equivalent role as the BHPS documentation web site
   `https://www.iser.essex.ac.uk/bhps/documentation/volb/index.html`.

2 Coverage and definition

The sample started in 1984, was extended to East Germany in 1990. The last available year is 2012. A rather stable set of core questions is asked every year covering the most essential areas of interest. Some of these areas are quite general to qualify the households and individuals. Some others areas are of particular interest for more specific topics. The following items are detailed in the DTC:

1. population and demography

2. education, training, and qualification

3. labour market and occupational dynamics

4. earnings, income and social security

5. housing

6. health

7. household production

8. basic orientation (preferences, values, etc.) and satisfaction with life in general and certain aspects of life.

Table 1: Some Special Topics Modules

Year	Wave	Sample	Topic
1992	I/9	A B C	Social security and poverty
1993	J/10	A B C	Further education or training
1994	K/11	A B C	Neighbourhood, values, and expectations
1995	L/12	A B C D	Use of time and preferences, increased range of income questions
1996	M/13	A B C D	Social network questions
1997	N/14	A B C D	Social security and poverty
1999	P/16	A B D C E	Neighbourhood, Values, Expectations
2001	R/18	A B D C E F	Social Networks, Working Conditions

Additionally, the basic information in one of these areas is enlarged by detailed questions every year as a special module. Consult Table 1.1, page 17 of the document. Of particular interest, we could find some examples given in the next Table: Among those special topics, van Praag and Ferrer-i-Carbonnel (2004) have used the *Income Evaluation Question* of 1994 to estimate a model of income subjective evaluation based on the lognormal distribution.

There are in fact two surveys. The SOEP which is the full version of the survey, available for researcu inside Germany and the GSOEP which covers 95% of the sample. The Department of Policy Analysis and Management at Cornell University provides an English version of the Public-Use file of the SOEP, the German Socio-Economic Panel (GSOEP), to researchers outside of Germany. For confidentiality reasons the GSOEP is a 95 percent sample of the full SOEP. It is this file, the GSOEP, that is used in the CNEF, the Cross-National Equivalent File, administered at Cornell University.

2.1 Samples

There are various samples in the data set sampled with different rates. All samples, except sample A are oversampled.

- A: German residents of West Germany. Sample A it is often called the West German Sample of SOEP. In 1984 it covered 4 528 households. A sampling probability of about 0.0002.
- B: Foreigners in West Germany. Turkish, Greek, Yugoslavian, Spanish or Italian household head. Sample B is oversampled and started with 1 393 households in 1984. The sampling probability was about 0.0008.
- C: Est German households. In June 1990, there were 2 179 households. A sampling probability of about 0.0004.
- D: Immigrants, started in 1995. The sampling probability is about 0.0002.
- E: Refreshment. In 1998, a new sample was selected from the population of private households in Germany. The sampling probability is about 0.00003.
- F: Innovation, new households added in 2000. Corresponds to 6 052 households, composed of 2 993 kids (age < 16) and 11 532 adult persons. The sampling probabilities are approximately 0.00028 for "German" households and 0.0005 for "non-German" households.
- G: High Income Sub-sample, households with a monthly income of at least DM 7 500 (EURO 3 835). Started in 2002. There were 1 224 households with 693 children (below 16) and 2 845 adults. Income limit raised to 4 500 euros later on.

2.2 Panel

In a panel survey, individuals and households are followed over the years. There is attrition because individuals can die, move, leave the household, quit the country. Individuals can also decide to stop participating. When an individual moves, he is followed, when he quit a household, a new household is created. Individuals can be lost for one year and come back into the survey. Because of attrition, new households were added.

Persons exit by:

1. Death
2. Moving abroad
3. Decide to stop participating

Persons enter by:

1. birth
2. moving into a SOEP household from somewhere else in Germany or from abroad
3. reaching age of 16 years (minimum respondents age)
4. new households and persons from a split of at least one old person from an old household

3 Data availability

The data can be obtained free of charge, but after a rather long accreditation process. A CD is sent by surface mail. A password is communicated on the phone, by a personal contact. A confidentiality agreement has to be signed.

3.1 What is on the CD ROM

An index.html file gives instruction on how to install the GSOEP. We have access to encrypted zip files. There are several formats: The strong advice we give is to use the Stata version. First, there should be a very handy module in Stata SE to access interactively the GSOEP. Second, if you do not have Stata or do not want to use it, you can access the files directly from `R`, using a routine that we shall detail below.

Software	English	German	English+German
Stata	-	-	`soep2012pw100.zip`
SPSS	`spss_de_v29.zip`	`spss_en_v29.zip`	
SAS	`sas_de_v29.zip`	`sas_en_v29.zip`	
PanelWhiz			`soep2012pw100.zip`
CSV			`csv_v29.zip`
CSV			`label_info_v29.zip`

3.2 Yearly files

There are two types of files. Either they are organized on a year basis, with the following list given in Table ltyear. We can note that they are files which are common to every year. Each year or wave is indicated by a letter from a (1984) to z (2009). There are files which are specific to one or two years. They correspond to specific domains which were explored for a given year.

The main files, which are common to all waves, are `p` and `h`. There are generated files from these two files `pgen` and `hgen`. The `pequiv` file will be described below.

Table 2: Yearly SOEP Codebooks and data files (February 2014)

File	Content	Codebook	Size
$P	Variables from personal questionnaires	`v29_p`	48.49 MB
$H	Variables from household questionnaires	`v29_h`	14.72 MB
$PGEN	Generated Person-Level Variables	`v29_pgen`	11.63 MB
$HGEN	Generated Household-Level Variables	`v29_hgen`	4.38 MB
$PKAL	Generated calendar variables from $P	`v29_pkal`	21.54 MB
$PEQUIV	Generated Person-Level variables for international comparison	`v29_pequiv`	20.38 MB
$KIND	Generated Child Variables (Person-Level)	`v29_kind`	2.82 MB

Documentation at: http://www.diw.de/en/diw_02.c.239921.en/codebooks.html

3.3 Variables names

Variable names are quite complicated, but obey some rules, the most important ones being their organization in digits. A correspondence table is as follows:

Digit	Meaning	Example
1	Wave (A, B, as in the BHPS)	the A in AP06
2	Unit of observation: P for personal and H for household	the H in AH27
3-4	Number attributed to the question	the 57 in AP57
5-6	Number of the item in the question	the 01 in AP3301

5 or 7	Indicating the sample specific question
	A = Auslander (sample B),
	O = Ostdeutcher (sample C)
		the last A in AP62A
		the letter O in HP42O

Very logic, but you must have the questionnaires! Nothing intuitive as in the BHPS. This naming is specific to the yearly files

3.4 Specific files

When exploring the list of yearly files, we get surprises:

Table 3: GSOEP file

Content	Prefix	File	Prefix	File	Prefix	File
	a	broad				
			p	biospe		
			p	br_exit		
			p	br_hhch		
			p	flege		
	a	h	p	h	z	h
	a	hbrutto	p	hbrutto	z	hbrutto
	a	hgen	p	hgen	z	hgen
			p	hrf		
	a	kind	p	kind	z	kind
	a	p	p	p	z	p
	a	phausl				
					z	page17
	a	pbrutto	p	pbrutto	z	pbrutto
	a	pequiv	p	pequiv	z	pequiv
			p	pfad		
	a	pgen	p	pgen	z	pgen
	a	pkal	p	pkal	z	pkal
	a	rtkalen				
			p	pluecke	z	pluecke
					z	vp
			p	wealth		

For each year (or wave) of SOEP data there are single data files for households (H) as well as for individual respondents (P) and children (KIND) based on interview information. The naming of the data files is wave-specific, starting with A for the first wave in 1984, B for 1985, ... , U for 2004.

1. The file `pkal` Employment and Income Calendar Files. Ever since its start in 1984, the SOEP contained a calendar section asking about employment status and sources of income received as of January through December of the previous year.

2. When individual temporarily drop out from the survey, some information can be rebuilt. This information is contained in the $PLUECKE file.

3. The wave-specific cross-section files $PBRUTTO encompass all individuals currently living in SOEP-households at a given point of time. These include respondents, children, and persons who refused to answer (unit-non-response).

4. ZPBRUTTO cumulates across all waves all individuals who left the household they lived in last year or even the survey.

5. The specific file `pflege`, `flege` for wave `P`, contains Persons Needing Care (Invalids).

6. BIOPAREN contains all individuals with at least one interview starting 1984. Using data from the biography questionnaire and the yearly information from the $P and $PGEN files, this file contains information on parents, which can be used for intergenerational analyses.

1. In order to facilitate the definition of longitudinal populations, the SOEP provides data files encompassing every individual respondent and child (file PPFAD) and any household (file HPFAD) ever contacted in the survey.

2. Cumulating drop-outs across all waves, YPBRUTTO contains information on the reason for temporary or permanent drop-out at the individual level.

3. Biography information at the individual level can be found in BIO.

4. Some information are detailed at the month level (different activities, income, etc). They are presumably in the `pkal` file.

5. Weights used to be in some specific files `PHRF` and `HHRF`. Try to find where they are now.

3.5 Long files

A panel is organized around two variables which are specific indicators that never change over the years. There is the household number and the individual number. These two indicators are a mean to built time series, repeated observations for the same variable. The list of files is given in Table 4. A panel is rather difficult to reconstruct. This is an advantage of having the long version of the panel. But some questions are asked for a given year. So it is not usefull or possible to reconstruct the panel for all the questions. Documentation is given at `https://data.soep.de/studies/1/datasets`

Table 4: GSOEP Long

File	Content	Size	Match
pl	Personal file long	1716	
hl	Household file long	316	
pgen	Person Generated	53	PERSNR
hgen	Household Generated	22	HHNR $HHNR
pkal	Person Calendar	475	PERSNR
pequiv	Cross-national Equivalent File	292	PERSNR
kidl	Data on children	10	
bio	Biography spell data visualized bioscope.exe (MSDOS)	77	
pbrutto	Person Gross File	44	PERSNR
hbrutto	Household Gross File	24	HHNR $HHNR
hpfad	Master Household File	0.3	HHNR HHNRAKT
hpfadl	Household Tracking File	12	
`pbr_exit`	Cumulated Exit	0.5	
ppfad	Master Person File	2	PERSNR
ppfadl	Individual Tracking File	74	
csamp	Sample Definition	0.4	

Documentation available at `https://data.soep.de/studies/1/datasets`.

The domain of Happiness Economics is supposed to be very important in the GSOEP. van Praag and Ferrer-i-Carbonell have built their book around the GSOEP. But it is difficult to track the variables that are present for all the waves. These variables are documented in the `pl` file.

- plh0182: Current life satisfaction
- plh0175: Satisfaction With Household Income

4 The `pequiv` file (and cnef)

Present both in the annual files and in the long term panel files, the pequiv file is particularly interesting. It is the *Cross-National Equivalent File*. It was created by Cornell University, in close cooperation with DIW-Berlin, ISER-Essex and StatsCan-Ottawa. It consists of variables from the German SOEP, American PSID, Canadian SLID and British BHPS, based on common definitions. The income variables are all annualized, meaning that the typical German SOEP variables asking about monthly income components have been transformed.

The Equivalent File variable names are identical across datasets, adding to ease of use. The reader is referred to the standard Equivalent File documentation in Burkhauser, Butrica, Daly, and Lillard (2001) to further information (all used original variables names from the data sets are included with the algorithms). The codebooks were available at Cornell University and are now at the Ohio State University `http://cnef.ehe.osu.edu/`, as found on the web page of Richard Burkhauser.

For ease of use, the German portion of the cross-national equivalent file has been included in the regular distribution of the SOEP data, both for the German and international distribution. In addition, the regular matching variable indicators HHNR, HHNRAKT, HHNR and PERSNR have been added (in addition to the already existing equivalent file matching variables such as X11101LL).

The German portion is found in the files PEQUIV, available starting 1984 (wave A) onward. The sampled population includes adult respondents, adult non-respondents and children in households with an interview, corresponding to the SOEP population.

Constructing post-government income in the GSOEP is a complicated task. The first task is to annualize income. Next an estimated tax burden for households or individuals must be computed using a tax estimation routine developed at DIW. This tax package produces estimated annual tax burdens for all households in the SOEP. These annual tax values are combined with the annualized components of income to create a measure of household post-government income.

Table 5: Some of the variables found in the PEQUIV file

Name	Content
HHNR	Original Household Number
PERSNR	Never Changing Person ID
X11101LL	Person Identification Number
D11102LL	Gender of Individual
D1110104	Age of Individual
D1110304	Race of HH Head
D1110404	Marital Status of Individual
D1110904	Number of Years of Education
D1111104	Satisfaction With Health
E1110104	Annual Work Hours of Individual
E1110204	Employment Status of Individual
I1110104	HH Pre-Government Income
I1110204	HH Post-Government Income
I1110504	HH Imputed Rent
I1110904	Total HH Taxes
M1110804	Have or had cancer
M1110904	Psychiatric problems
M1112004	Health limits kneeling
M1112104	Health limits vigorous activities
M1112204	Body height
M1112304	Body weight

5 Questionnaires

Using questionnaires can help understanding how the variables are coded and in which file of the corresponding wave they can be found. The *long* dataset also follows those rules. Be careful: some modules were introduced for some years only. The following description is not about those special modules.

Questionnaires for each year can be found on the SOEP website at

`http://www.diw.de/en/diw\_02.c.222729.en/questionnaires.html`

The questionnaires cover the whole surveyed period but the English translation is only provided starting from the year 2000. The central questions of the survey can be found either in the individual or in the household questionnaires.

5.1 The Individual Questionnaire

The answers are stored in the `$p` file, the `$pkal` contains the answers to section 4. In the long survey: most of the information is reported in the `pl` file, the answers of section 8 are in the `bio` file.

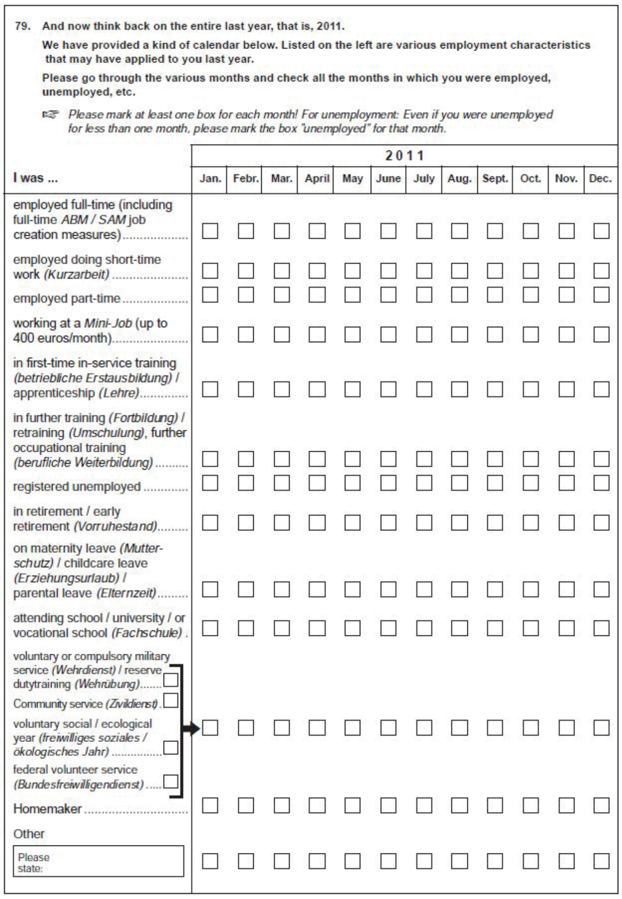
1. **How is it done?** *Introduction on how to answer the questions*

2. **Your current life situation.** *It contains questions about satisfaction with life, leisure and working hours, main activity of the respondent: student, involved in some training or looking for a job. This last part had a different section in some older questionnaires: only for those not employed.*

3. **Your current job+some more questions for everyone.** *Important information about their current job (weekly or monthly based), here you can find questions about hours worked, overtime, earnings (both gross and net) of last month, informations about the type of job, occupation status, whether the respondent changed job last year, durations of the employment length, second job and other benefits (pensions, unemployment, etc..).*

4. **How were things last year.** *Be careful here (also true for the next section): all the information of this section refers to the previous year! i.e if the survey year is 2010 then the questions will be about the previous year (2009). About education last year, job and a useful calendar that allows to reconstruct the full year of the respondent, see Figure*

1. eg. if the respondent was self-employed from July to August and unemployed the rest of the year.



79. And now think back on the entire last year, that is, 2011.

We have provided a kind of calendar below. Listed on the left are various employment characteristics that may have applied to you last year.

Please go through the various months and check all the months in which you were employed, unemployed, etc.

☞ *Please mark at least one box for each month! For unemployment: Even if you were unemployed for less than one month, please mark the box "unemployed" for that month.*

	2011											
I was ...	**Jan.**	**Febr.**	**Mar.**	**April**	**May**	**June**	**July**	**Aug.**	**Sept.**	**Oct.**	**Nov.**	**Dec.**
employed full-time (including full-time *ABM / SAM* job creation measures)	□	□	□	□	□	□	□	□	□	□	□	□
employed doing short-time work *(Kurzarbeit)*	□	□	□	□	□	□	□	□	□	□	□	□
employed part-time	□	□	□	□	□	□	□	□	□	□	□	□
working at a *Mini-Job* (up to 400 euros/month)	□	□	□	□	□	□	□	□	□	□	□	□
in first-time in-service training *(betriebliche Erstausbildung)* / apprenticeship *(Lehre)*	□	□	□	□	□	□	□	□	□	□	□	□
in further training *(Fortbildung)* / retraining *(Umschulung)*, further occupational training *(berufliche Weiterbildung)*	□	□	□	□	□	□	□	□	□	□	□	□
registered unemployed	□	□	□	□	□	□	□	□	□	□	□	□
in retirement / early retirement *(Vorruhestand)*	□	□	□	□	□	□	□	□	□	□	□	□
on maternity leave *(Mutterschutz)* / childcare leave *(Erziehungsurlaub)* / parental leave *(Elternzeit)*	□	□	□	□	□	□	□	□	□	□	□	□
attending school / university / or vocational school *(Fachschule)*	□	□	□	□	□	□	□	□	□	□	□	□
voluntary or compulsory military service *(Wehrdienst)* / reserve dutytraining *(Wehrübung)* □ Community service *(Zivildienst)* □ voluntary social / ecological year *(freiwilliges soziales / ökologisches Jahr)* □ federal volunteer service *(Bundesfreiwilligendienst)* □	□	□	□	□	□	□	□	□	□	□	□	□
Homemaker	□	□	□	□	□	□	□	□	□	□	□	□
Other Please state:	□	□	□	□	□	□	□	□	□	□	□	□

Figure 1: Sample of Information included in $pkal

5. **Income last year.**

   The amounts of the sources of income of the previous year, if the respondent received extra pay from employment, some additional questions for pensioners.

6. **Health and Illness**

   Information about health insurance and self-assessed health conditions.

7. **Attitudes and opinions**

   Mainly politics and concerns about some aspects of the society in general.

8. **Family situation and background (sometimes called: Citizenship and origins)**

   Sex, birth month, birth year, marital status, relationship, citizenship, origins, risk-adversity, satisfaction with life in general and what changed last year in terms of important events (having a child etc..).

9. **Implementation of the interview**. *The interviewer fills in the informations about the length of the interview, date etc..*

5.2 The Household Questionnaire

The answers are stored in the `$h` file. In the *long*: most of the informations are reported on the `hl`.

1. **How is it done?**

   Introduction on how to answer the questions

2. **The costs of your dwelling (sometimes: expenses for household and flat), subsection for tenants and homeowners+ more questions**

   Informations about the house, the neighborhood, the landlord, how the respondent got it, expenses for loans, housekeeping and renovations. Also, whether the respondent has commodities, such as: microwave, cars etc.. and if they were bought last year.

3. **Does your household have..?**

   This section asks questions about assistance and children: where they were born, if they go to school and the costs of it for the household.

4. **Implementation of the interview**

   The interviewer fills in the pieces of information about the length of the interview, date etc..

Other questionnaires can be found in different years (mainly starting from 2003), they are all supplement to the individual questionnaire:

- Youth Questionnaire (english in 2001)
- Supplementary Biography Questionnaire (english in 2001)
- Short Questionnaire “Luecke”)

- Mother and Child Questionnaire (newborn) (english in 2012)
- Mother and Child Questionnaire (2-3yearolds) (english in 2012)
- Mother and Child Questionnaire (5-6yearolds) (english in 2012)
- Parents Questionnaire (7-8yearolds)
- Mother and Child Questionnaire (9-10yearolds) (english in 2012)
- Pupil Questionnaire for the 11-12 years old
- Questionnaire “The deceased person”

6 Extracting data

Chapter 4 in the big documentation *Desktop Companion to the German Socio-Economic Panel (SOEP)*. Data are organized in several files,depending on the level of analysis. In the yearly version, there is a separate file for each year (as in the BHPS). In the long version, all the years were regrouped in big files by domain. The data can be matched using identifiers

1. PERSNR is unique over time and individuals. It is the PID in the BHPS.
2. UHHNR is the current household identifier. Households have not a unique number, because they can be created, split, etc..

In this trivial example, we are interested in information from 1984 and 1985 at the person and household level. Below is a listing of those variables mentioned.

```
+---------+--------+--------+-------------------------------------------+
| VARNAME | FILE | YEAR | LABEL |
+---------+--------+--------+-------------------------------------------+
| AP06 | AP | 1984 | School-Leaving Degree |
| BP16 | BP | 1985 | Employment Status |
| AH02 | AH | 1984 | Change in HH comp Since Jan 1st Prev yr |
| BH01 | BH | 1985 | Children under age 16 in HH |
+---------+--------+--------+-------------------------------------------+
```

6.1 R Codes

The idea of this section is to provide some basics data-cleaning commands, using the language R and files in the STATA format (files ended with .DTA). The first step is to install the library "foreign" on your computer. It includes commands to open files of a great variety of formats.

```
install.packages("foreign")
```

Then, for reading STATA files, we have the following instructions:

```
library("foreign")
data <- read.dta("C:/Lara/Data/GER/2009/zp.dta")
```

Here we are working with the individual file p of wave z (2009).

Remark:

> This file is huge. The variables of interest can be browsed on the website:
>
> https://data.soep.de/search/variables?page=1
>
> Let's say we are interested in the questions of section 2 of the individual questionnaire. In particular, we want to know the respondent gross income last month and hours per week (actual) that have been worked. A way to restrict the research options on the website is to select
>
> SOEP Core Study
>
> on the right side of the website under "Study", *Individual level* under "Analysis unit", and 2009 under "Period". We find that the variable for gross income last month is zp7201 and the one for actual weekly hours actual is zp62.

We create therefore a new database with those variables plus the household and individual identifiers of the respondents IN ORDER TO BE ABLE TO MERGE DIFFERENT FILES.

```
varsofint <- c("hhnr","persnr","zp7201","zp62")
data.sel <- data[varsofint]  # Extracting variables
                             # from a dataset
                             # using variable names
names(data.sel) <- c("hid","pid","Ymonth","Hweek")
                             # changing the names in
                             # the new dataset
```

With the last line we changed the names of the variables of the new dataset, in order to have something easier to remember. Now, let's say we want to add household and individual weights to the database, the variables are called respectively `w1110209` and `w1110509` and they are in the database `zpequiv`. Therefore we need to merge the two data bases. This a very common and recurrent operation when manipulating those files. For this, we need the household and the individual identifiers.

Remark:

> Be careful: if you use the *long* version of the GSOEP you need to merge ALSO using the variable *syear* (survey year.)

So reading the needed weights:

```
data.weight <- read.dta("C:/Lara/Data/GER/2010/zpequiv.dta")
varsofint <- c("hhnr","persnr","w1110209","w1110509")
data.weight.sel <- data.weight[varsofint]
names(data.weight.sel) <- c("hid","pid","hhweight","indweight")
```

And now merging the two files `data.sel` and `data.weight.sel`, using the variables `hid` and `pid` for merging.

```
data.fin <- merge(data.sel,data.weight.sel,by=c("hid","pid"),all=F)
data.fin.noNA <- na.omit(data.fin)
```

The last line removes the `NA` from the database. Note that this might not be necessary. There are operations in `R` that can be done with `NA` observations. You just have to introduce a specific option. For instance

```
mean(x, na.rm = TRUE)
```

Now we inspect the variables to see whether they have weird values or the coding has a special meaning, using the `summary` command.

```
data.toclean <- data.fin.noNA
summary(data.toclean$Ymonth)
summary(data.toclean$Hweek)
```

The first line only changes the name of the database, the second and the third give us the statistics of the variables `Ymonth` and `Hweek`. We get the following results:

```
> summary(data.toclean$Ymonth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     -3      -2      -1    1248    2140   50000
> summary(data.toclean$Hweek)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     -3      -2     150     209     400     999
```

Those results are weird, we come back to the website to check the meaning of the weird values, what we find is that: -3 stands for invalid, -2 for not applicable and -1 for no answer. What we have to do is therefore removing the values of both Ymonth and Hweek that are smaller than zero.

```
id = data.toclean$Ymonth>=0 & data.toclean$Hweek>=0
data.cleaned <- data.toclean[id,]
```

Other useful checks to make sure that your dataset is well cleaned:

1. Same measurement units for all the variables. eg. weekly hours with weekly wage.

2. Check for censoring that may come both from the way the question is asked or from the way the variable is coded.

3. Define thresholds (or criteria) for outliers and remove them.

4. Make the dataset coherent with economics. i.e. no observations below the minimum wage.

5. Whatever decision you make be sure you can defend it and always do robustness checks.

Now we have a database that is cleaned, we may be interested, especially when working with a lot of different datasets, in cleaning the console and only keep this last object. One way to do it is:

```
rm(list=setdiff(ls(), "data.cleaned"))
```

We can also export the database in an excel file so that, the next time, we can work directly with the database that we just cleaned.

```
setwd("C:/Lara/Data/GER/Workshop")
write.csv(data.cleaned, file="data.csv")
```

The first line indicates the path to the folder where you want to store the excel database. To open it again then we only need to:

```
setwd("C:/Lara/Data/GER/Workshop")
data <- read.csv("data.csv", sep=",", header=T)
```

The dataset is now ready,
you can start working,
have fun!