dal 1994 Consorzio Interuniversitario ALMALAUREA

“MOnitoring Trends In Vietnamese graduates’ Employment”

MOTIVE www.motive-euproject.net (Project No. 609781-EPP-1-2019-IT-EPPKA2-CBHE-SP)

Short intensive training *on data cleaning: AlmaLaurea experience*

Silvia Galeazzi, Claudia Girotti (AlmaLaurea)

Thai Nguyen city, 26-29 October 2021

Co-funded by the Erasmus+ Programme of the European Union

Suggestions for first MOTIVE Employment Status Survey

At the beginning of a survey common to all universities, is better to:

- **contact all graduates for which is intended to run the survey** (in this case 2020 graduates, 1 year from graduation)
- **carefully define the questionnaire**
- **bet on quality of collected data**, so to have as output a good comparative report of all 9 universities involved
- focus on team work and collaboration between universities

Record layout and identification of valid respondents

If in the database of the interviewees there are graduates not included in the initial list, it is necessary to verify if the problem is due to:

- **database merge error:** an error occurred during database merging → repeat the merge

- **incomplete population database:** the initial list of graduates used in this comparison is not complete → repeat the merge using the correct databases

- **error in the starting list:** the list of graduates for the CATI survey contained incorrect names that did not correspond to the population to be interviewed. For example, the list contains not only graduates but also undergraduates → eliminate interviews referred to people who are not part of target population

In Vietnam, each university proceeds through their offices in charge of telephone contact with graduates (some important points: homogeneity of survey period, use of the same questionnaire, use the same way of formulating questions, …)

Record layout and identification of valid respondents

All collected questionnaires need to be checked.

It is necessary to verify that the database record layout is **complete** with all the variables foreseen by the questionnaire (**alphanumeric name** is preferable).

The **identification code** of each graduate must be present, so…

Administrative data + Graduates’ Employment survey

It is very important to verify that the database contains **only interviews of graduates actually involved in the survey**. A comparison must be made between the initial list of graduates and the list of graduates interviewed.

Duplicate interviews

It is very important to verify that the database does not contain **double or multiple interviews referring to the same graduate**.

If this occurs, it is necessary to compare the two interviews:

- if they are the **same**, then one of the two records must be deleted;
- if they are **different**, it is necessary to evaluate which of the two interviews must be considered valid (for example based on criteria of completeness of the interview or consistency between the answers given).

Completeness of interviews (1 of 4)

Check that **the interview is complete** for each graduate interviewed.

If an interview has been **interrupted**, it is necessary to assess whether:

- it can be considered **valid** (for example, if it is complete at least x% of the questions);

- it must be **eliminated** (for example if the information collected is few and not usable).

Completeness of interviews (2 of 4)

In the case of an interrupted interview, it is necessary to evaluate the **completeness percentage** (the number of completed questions).

Alternatively, we can define "**fundamental questions**": in this case we consider as valid an interview in which the fundamental questions are completed, regardless of the completeness of the questionnaire.

It is essential to decide together which criteria to use.

Completeness of interviews (3 of 4)

It is necessary to make a descriptive analysis of the number of complete interviews and interrupted interviews: through the simple frequency of each question we can evaluate the presence of **missing values**, which are different from the **no-answers** (some people decided to take part in the survey but not to answer a specific question).

Missing values are allowed only in accordance with the filters and skips required by the questionnaire: for example, if a graduate does not work, the section related to the characteristics of the work performed must contain missing values.

Completeness of interviews (4 of 4)

More team work and more collaboration between universities: communicate more each other. The team research leader has the important role of coordinating all the researchers**. It is important that you use the same method** (for example if one university applies the criterion of "fundamental questions", also the other universities have to apply it). It is useful for you collaborate, because for example, if you have some doubts on application of the criterion, you could compare yourself and ask the doubts common to AlmaLaurea staff.

Variable labels and value labels

Define the labels of each variable and of each value in the database: each variable must have a "**talking" label**” (short description of the topic), according to what is indicated in the survey questionnaire.

→ **The information contained in the database must be clear to the researcher.** This allows a clear and appropriate use of the database, even after some time and by other researchers.

Recoding of open answers

For each question in the questionnaire that includes the answer “**other, specify**” it is necessary to create a new variable that has the same name as the original variable but ending with “_rec”: ***Q7 “You have answered that you are not looking for a job; why aren’t you looking for it?"*** it is necessary to create a new variable Q7_rec.

	Q7	Q7_string	Q7_rec
id1	1		
id2	3		
id3	8	I became a billionaire	6
id4	8	…	no answer

Q7. You answered that you are not looking a job; why aren't you looking for it ?
[01] continuing studies/further training after graduation
[02] voluntary civil service
[03] waiting to be called back after having passed a test/competitive examination/selection or waiting to start a self-employment-activity
[04] opening own activity (entrepreneur)
[05] having a break for self-orientation (gap year)
[06] personal reasons (homemaker, maternity leave, looking after children or relatives, health reasons, retirement, etc.)
[07] no job opportunities
[08] other reason [SPECIFY]

code 8 → code 6, that is «*personal reasons*»

If the string «I became a billionaire»
is not recodable in one of the other categories
and a lot of people write it down in the string,
we could think to add this category of answer
for the next survey,
during the revision of the questionnaire

Notes to the interview

It is important read the "**notes to the interview**" field, if present.

This field can contain **important information and clarifications useful for defining the graduate's employment situation**: they can therefore be useful for verifying the correctness of the answers included in the previous questions.

The field "notes to the interview" can also contain updated information regarding the **telephone number or e-mail address.** It is important to enter this information in a specific field. In this way we can have updated contact details for future contacts.

Checking the path and filters of the questionnaire

- **check the codes and their labels**: there must be no codes or labels other than those provided for in the questionnaire. This check can be made through the frequency of each variable.

- **weight of the no answers**: if in a variable the number of no answer is too high, it is important to evaluate the quality of the information collected through that question. If the researcher believes that the quality of the information is poor, it is better not to use the variable in the analysis.

Anyway, if the number of no answers is high, it is necessary to evaluate whether to improve the formulation of the question or the formulation and completeness of the answer modalities for future surveys.

- **check the path of all the interviews**: to verify that the respondents have answered all and only the relevant questions.

Checking the path and filters of the questionnaire

- Checking the path of the interviews must be carried out first of all through the **frequency of each variable** to verify the number of respondents to each question.

- Secondly, we must **cross between two variables.** If, by mistake, a graduate has answered a question he did not have to answer, then we must delete the answer to that question (delete the value in the variable)

For example:

64 interviews, 2 questions analyzed:

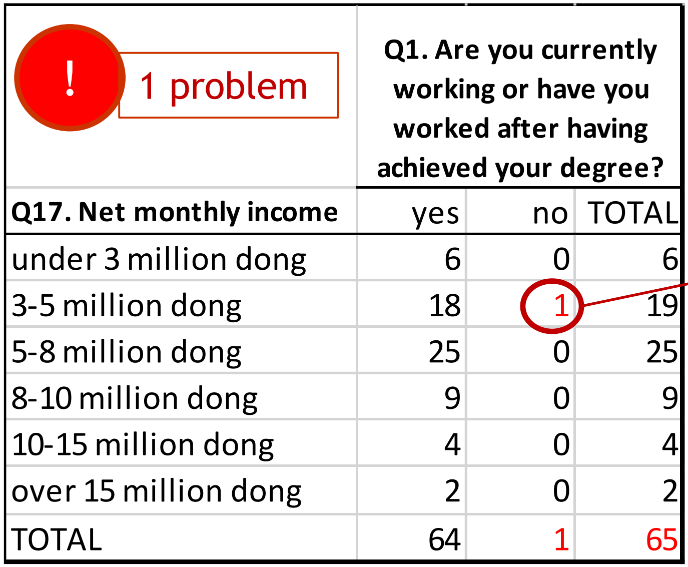
- Q1 *"Are you currently working or have you worked after having achieved your degree?"*

- *Q17 "Net monthly income"*

OK

	Q1. Are you currently working or have you worked after having achieved your degree?		
Q17. Net monthly income	yes	no	TOTAL
under 3 million dong	6	0	6
3-5 million dong	18	0	18
5-8 million dong	25	0	25
8-10 million dong	9	0	9
10-15 million dong	4	0	4
over 15 million dong	2	0	2
TOTAL	64	0	64

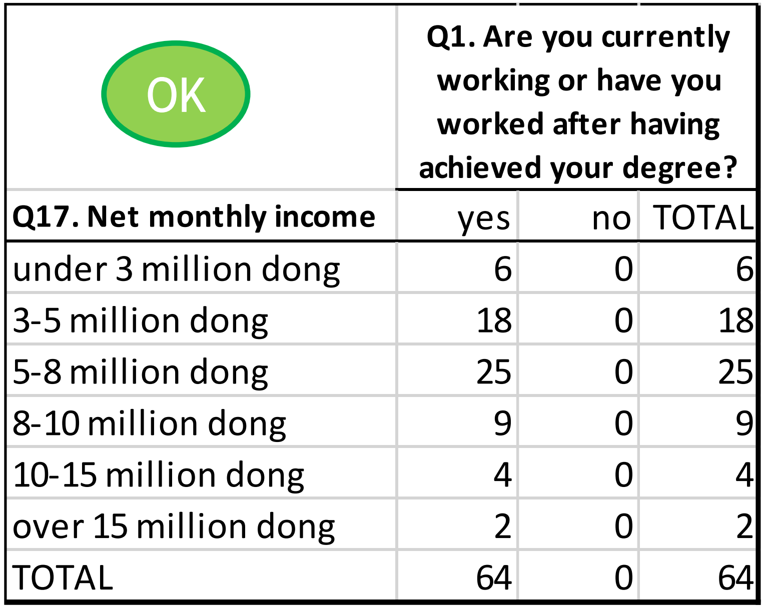
Checking the path and filters of the questionnaire: example



! 1 problem

	Q1. Are you currently working or have you worked after having achieved your degree?		
Q17. Net monthly income	yes	no	TOTAL
under 3 million dong	6	0	6
3-5 million dong	18	1	19
5-8 million dong	25	0	25
8-10 million dong	9	0	9
10-15 million dong	4	0	4
over 15 million dong	2	0	2
TOTAL	64	1	65

It should be deleted,
it is an error



OK

	Q1. Are you currently working or have you worked after having achieved your degree?		
Q17. Net monthly income	yes	no	TOTAL
under 3 million dong	6	0	6
3-5 million dong	18	0	18
5-8 million dong	25	0	25
8-10 million dong	9	0	9
10-15 million dong	4	0	4
over 15 million dong	2	0	2
TOTAL	64	0	64

Checking the path and filters of the questionnaire: example

!

2 problems

	Q1. Are you currently working or have you worked after having achieved your degree?		
Q17. Net monthly income	yes	no	TOTAL
under 3 million dong	5	0	5
3-5 million dong	18	1	19
5-8 million dong	25	0	25
8-10 million dong	9	0	9
10-15 million dong	4	0	4
over 15 million dong	2	0	2
TOTAL	63	1	64

It should be deleted,
it is an error

We shoud have a total of 64, while we have 63
→ we should add 1 case of not answer in Q1

OK

	Q1. Are you currently working or have you worked after having achieved your degree?		
Q17. Net monthly income	yes	no	TOTAL
under 3 million dong	5	0	5
3-5 million dong	18	0	18
5-8 million dong	25	0	25
8-10 million dong	9	0	9
10-15 million dong	4	0	4
over 15 million dong	2	0	2
not answer	1	0	1
TOTAL	64	0	64

Consistency checks

It is possible to compare the **consistency between the answers** given to different questions: if the answers are conflicting, it is necessary to understand exactly the real situation of the interviewee and then modify one of the two answers in the database.

Creation of new variables

In some cases it might be useful to create **new variables** to make the information more explanatory.

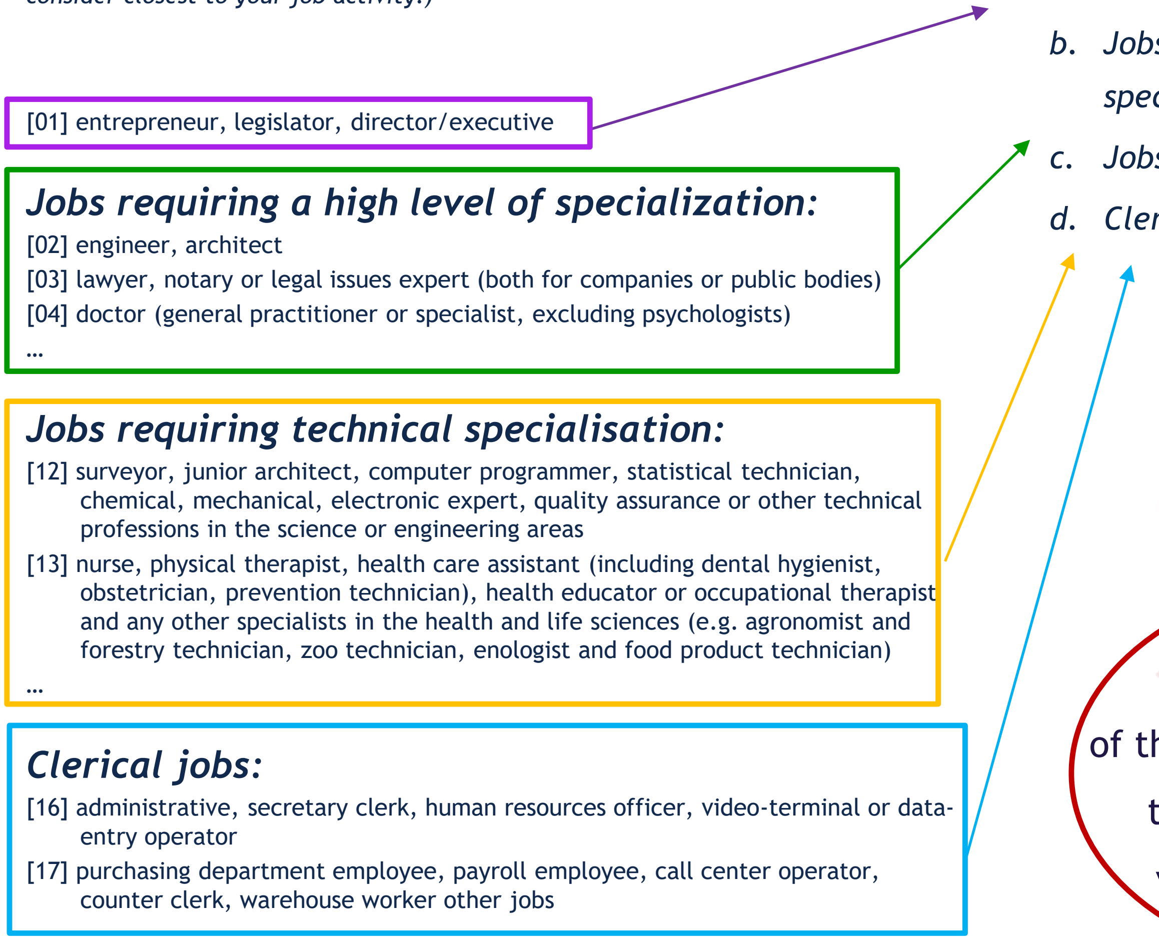
→ aggregation of items of the same variable

→ combination of two or more variables

Creation of new variables: example of aggregation

Q11. *"What is your current job?"*

(If you are performing different job activities, the answer should refer to the prevalent job, according to any criteria; the list below enumerates different jobs, based on the area and level of specialization. You should select only the one you consider closest to your job activity.)



[01] entrepreneur, legislator, director/executive

Jobs requiring a high level of specialization:

[02] engineer, architect

[03] lawyer, notary or legal issues expert (both for companies or public bodies)

[04] doctor (general practitioner or specialist, excluding psychologists)

…

Jobs requiring technical specialisation:

[12] surveyor, junior architect, computer programmer, statistical technician, chemical, mechanical, electronic expert, quality assurance or other technical professions in the science or engineering areas

[13] nurse, physical therapist, health care assistant (including dental hygienist, obstetrician, prevention technician), health educator or occupational therapist and any other specialists in the health and life sciences (e.g. agronomist and forestry technician, zoo technician, enologist and food product technician)

…

Clerical jobs:

[16] administrative, secretary clerk, human resources officer, video-terminal or data-entry operator

[17] purchasing department employee, payroll employee, call center operator, counter clerk, warehouse worker other jobs

Q11_AGGR

a. Entrepreneur, legislator, director/executive

b. Jobs requiring a high level of specialization

c. Jobs requiring technical specialisation

d. Clerical jobs

For the homogeneity of the Report, it is fundamental to define **all together** which variables to create and how

Creation of new variables

In some cases it is necessary to **exclude some values from the analysis**, so as to consider only the valid cases (for example in the case of calculating the average). This is possible by defining missing values or using filters or creating new variables.

For example,

in question Q17 ***"What is your net monthly income?"***

(Remember that your answers are protected by the legislation on the protection of personal data and that they will be used for no reason other than statistical purposes). (If you have more than one job, refer to the main one)

[01] under 3 million dong

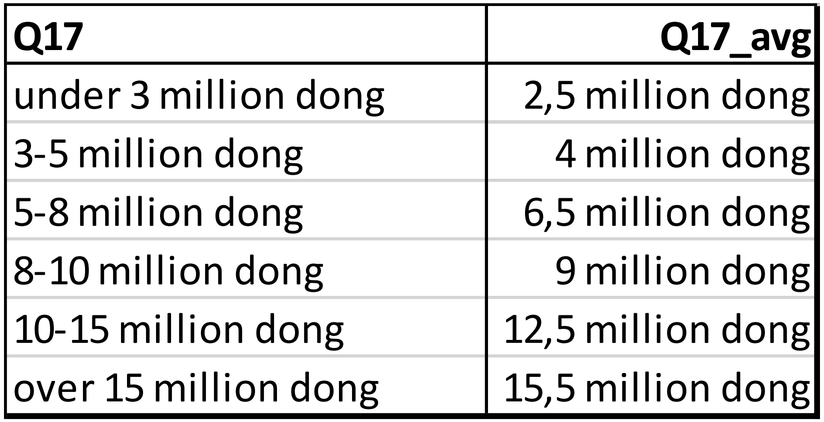
[02] 3-5 million dong
[03] 5-8 million dong
[04] 8-10 million dong
[05] 10-15 million dong
[06] over 15 million dong

there are several earning ranges

Creation of new variables

The average must be calculated by excluding no-answers and by using the central value of each earning range (except the first one, i.e. under 3 million dong, and the last one, i.e. over 15 million dong).

We can create a new variable named Q17_avg, which takes these values:



Q17	Q17_avg
under 3 million dong	2,5 million dong
3-5 million dong	4 million dong
5-8 million dong	6,5 million dong
8-10 million dong	9 million dong
10-15 million dong	12,5 million dong
over 15 million dong	15,5 million dong

If a graduate has not answered the question Q17 → Q17_avg = no value.

In this way, to calculate the average salary, it is more convenient to use the variable Q17_avg than the variable Q17.

Merge between administrative data and interviews

Administrative data + Graduates’ Employment survey

identification code of the graduate

It is useful to create a variable ("**interv**") which takes values:

0 “non-interviewees”

1 “interviewees”

The "interv" variable must have value 1 only for the interviews that have "passed" the checks described in the previous points. For interviews interrupted or eliminated because of poor quality, "interv" has a value 0.

→ The frequency of this variable provides the **response rate**.

Merge between administrative data and interviews

interv = 1 → to limit the analysis to only graduates who answered the questionnaire

interv = 0 → to analyze the structure of the collective of non-respondents and compare it with that of the target population, in order to highlight any bias.

Population target

Collective of respondents

It is important that all these database cleaning operations are carried out following the same procedures, through a perfect team work.

Next meetings

During next meetings, suggestions for …

- **“data visualization”** (charts, graphs, …)
- **written Report** (structure of the Report, chapters, conclusions, …)
- **methodological notes** for the online visualization of data

silvia.galeazzi@almalaurea.it

claudia.girotti@almalaurea.it