AdTrans 2023-1-PL01-KA220-HED-000158917

ADTRANS
ADVANCED DATA ANALYSIS FOR TRANSPORTATION AND MOBILITY ACADEMIC TEACHERS

Introduction to statistical analysis with SPSS

Ing. Mgr. Hana Brůhová Foltýnová, PhD.

Co-funded by the European Union

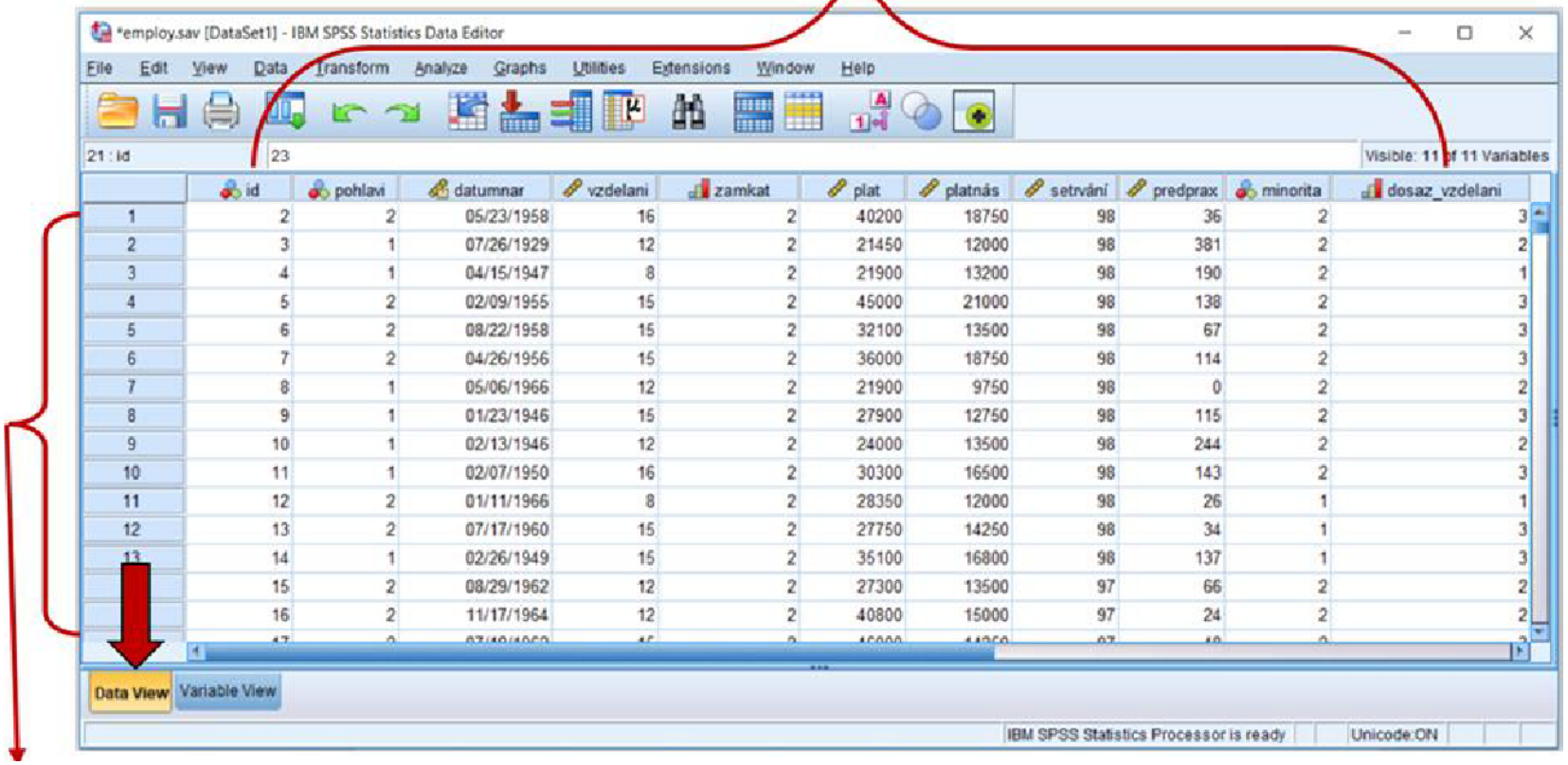
DATA ANALYSIS SOFTWARE

- The IBM SPSS statistical software consists of two basic windows and two types of additional windows. The basic windows represent a separate view of the data - that is, the respondents' answers (the so-called data window) and a separate view of the outputs resulting from the analyses run (the so-called output window). Additional windows then serve to specify the requirement before running the selected analysis (syntax window or script window). Let's now introduce the individual windows of this program in more detail.
- Individual window types differ in their function, display, and menu within their "Menu". However, all windows are systemically interconnected, even though they form separate units that can also be saved separately.
- IBM SPSS software allows you to open multiple windows at once (of the same or different types), between which you can switch at will. This allows you to work with multiple input data at once and determine which file you are currently analyzing. The file that is currently activated for analysis is marked with a red cross in the icon of the respective toolbar. If you want to activate another window, click on its icon with the mouse.

SPSS – “Data View” sheet

- Data View Sheet
- allows you to view the (collected and already encoded) data itself
- format of a data matrix that contains rows and columns
- rows = individual "cases", which we can most often imagine as individual persons (respondents) who participated in our research (however, research units can also be groups of people, various territorial units, individual types of social service organizations, etc.)
- columns = individual "variables", or properties that relate to these cases (to the surveyed persons - respondents), to which icons symbolizing the method of measurement (measuring scale) of the given variable are also assigned
- the number of columns or rows in the data matrix is not limited in the data window, the "Data View" sheet allows us to view or edit the data

Columns = individual variables = questionnaire questions"



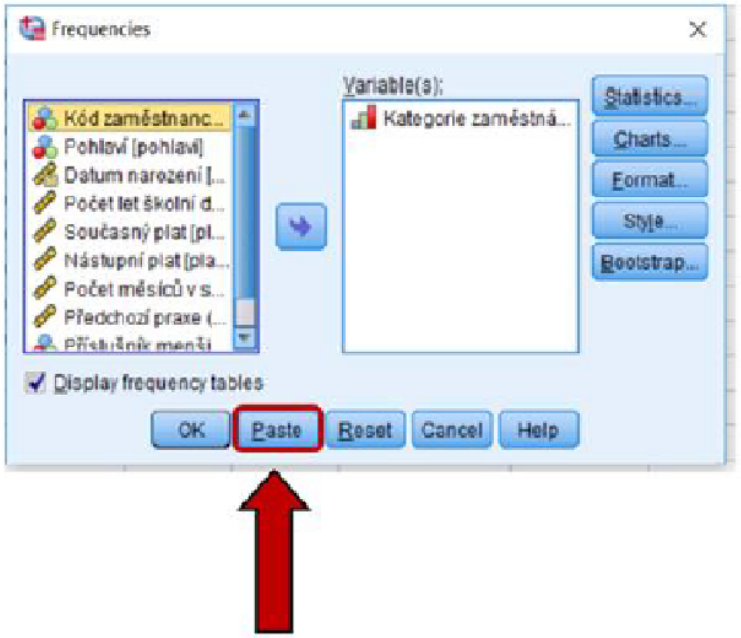
	id	pohlavi	datumnar	vzdelani	zamkat	plat	platnás	setrvání	predprax	minorita	dosaz_vzdelani
1	2	2	05/23/1958	16	2	40200	18750	98	36	2	3
2	3	1	07/26/1929	12	2	21450	12000	98	381	2	2
3	4	1	04/15/1947	8	2	21900	13200	98	190	2	1
4	5	2	02/09/1955	15	2	45000	21000	98	138	2	3
5	6	2	08/22/1958	15	2	32100	13500	98	67	2	3
6	7	2	04/26/1956	15	2	36000	18750	98	114	2	3
7	8	1	05/06/1966	12	2	21900	9750	98	0	2	2
8	9	1	01/23/1946	15	2	27900	12750	98	115	2	3
9	10	1	02/13/1946	12	2	24000	13500	98	244	2	2
10	11	1	02/07/1950	16	2	30300	16500	98	143	2	3
11	12	2	01/11/1966	8	2	28350	12000	98	26	1	1
12	13	2	07/17/1960	15	2	27750	14250	98	34	1	3
13	14	1	02/26/1949	15	2	35100	16800	98	137	1	3
	15	2	08/29/1962	12	2	27300	13500	97	66	2	2
	16	2	11/17/1964	12	2	40800	15000	97	24	2	2

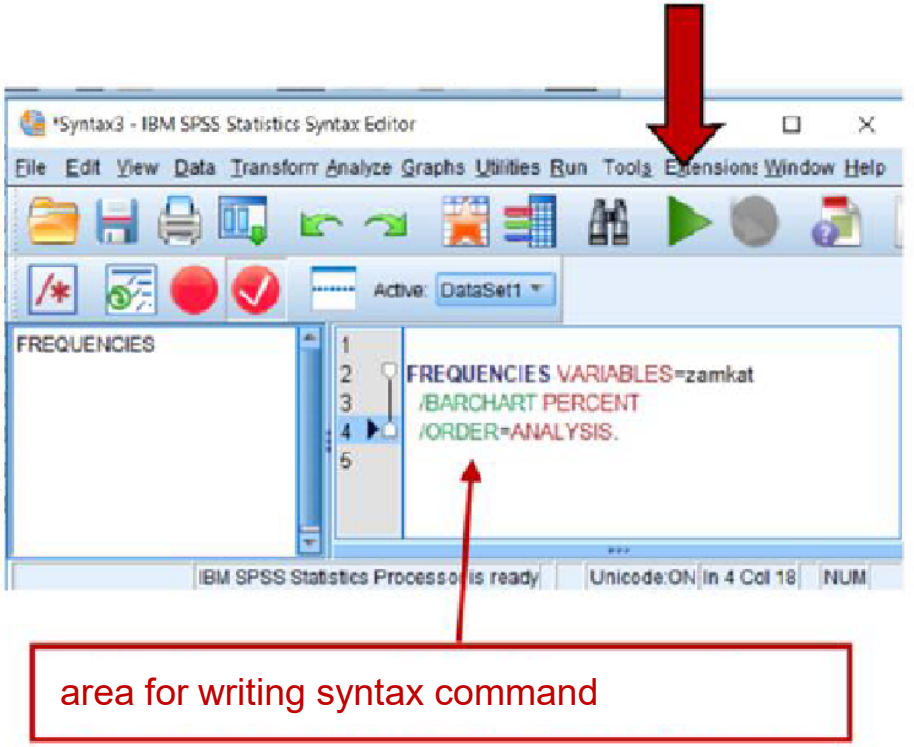
Rows = individual variables = answers"

Syntax window – Syntax

- used to enter commands leading to the launch of individual analyses
- It uses commands in the form of sentences that have a predetermined structure, instead of us entering the necessary command in the "Menu".
- SPSS can therefore be controlled in two ways: either using the menu bar in the data or output windows, or using a special language designed for entering commands in this software - syntax.
- The syntax command can be saved for future use as a separate file (with the extension ".sps"), which can be recalled at any time (without having to type the entire command again). It can also be sent to other users, for example, for the purpose of using the same sequence of analyses that would be required together.
- For each task that we enter into the IBM SPSS program in a standard way for processing (i.e. using the menu bar options from the data or output window), there is a syntax sentence that can also be used to call this task. If we want to know the syntax command sentence for a given procedure, or save it for future use, we set the "display syntax sentence" function in the "Menu".
- The syntax sentence corresponding to the given operation is then part of the output - it can always be found on the first lines of the Output.
- The syntax window opens automatically in IBM SPSS only if this option is selected in the program settings preferences. If this function is not preset a priori, a new syntax window can be called up using the menu command: FILE → NEW → SYNTAX, or via the "Paste" button in the analysis dialog boxes.
- In the right window, enter the command of the operation you want to perform and click the box with the green arrow in the menu bar of the syntax window (this will start the procedure). However, we will not use the syntax window for this study material.

Syntax window – Syntax

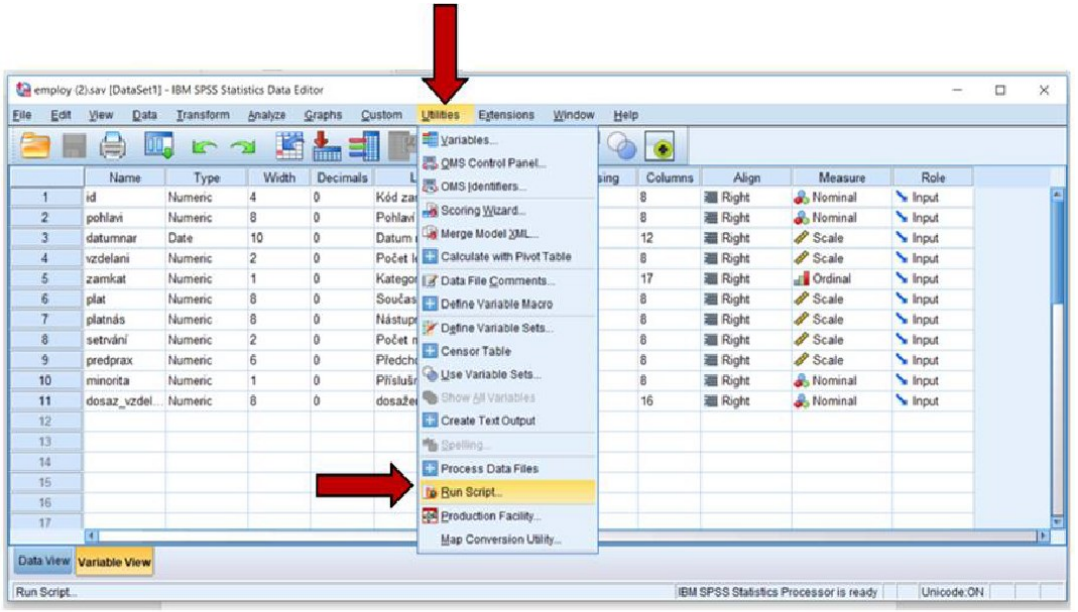


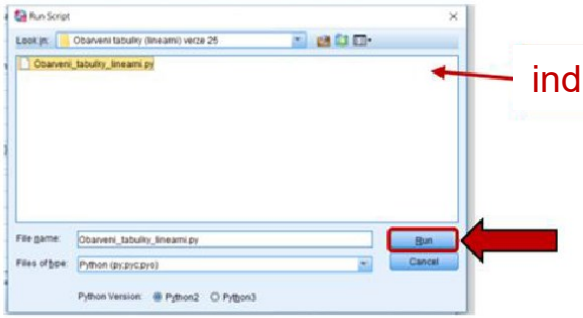


Script window - "Script"

- The Script window – called “Script” – is used to write and run scripts (short programs) that can automate certain tasks in IBM SPSS. Using the selected script, we can manipulate our outputs – we can edit them, run data transformations or selected statistical procedures, export graphs, color or sort tables, etc.
- There are a number of scripts for IBM SPSS. Some of them are included in the software installation, others can be downloaded from the IBM website. Each user also has the option of creating their own script programs (with basic programming knowledge) or modifying existing scripts as needed.
- The contents of the script window can be saved as a separate file (with the extension ".sbs") for future use.

Script window - "Script"



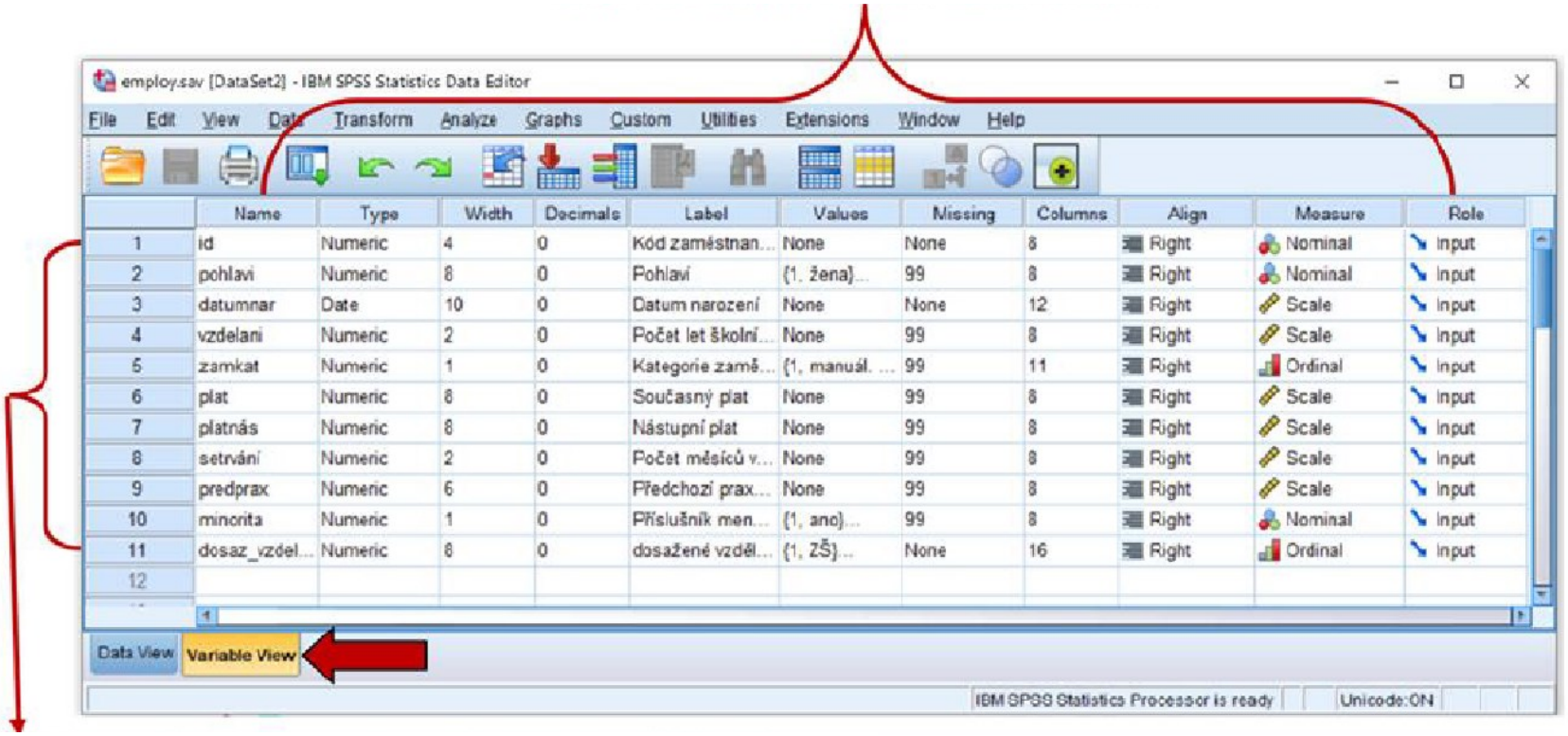


indication of places where the script is used

"Variable View" sheet – view of variables

- allows a detailed view of individual variables, or rather their specific descriptive characteristics
- the description of variables in the "Variable View" sheet must **always precede** the actual entry of data from the questionnaire survey into the data matrix
- **rows** in the data matrix, in the "Variable View" sheet represent **individual variables** , or individual questions from the questionnaire (unlike the "Data View" sheet, where variables are placed in columns!!!) - the number of rows is not limited in any way
- **columns** in the data matrix, in the "Variable View" sheet, indicate **the basic characteristics** of individual **variables** (such as variable name, variable label, etc. - more details later in this chapter) - there are always exactly 11 columns

columns = characteristics of variables



rows = individual variables = questionnaire questions

Definition of individual variables – “Variable View” sheet

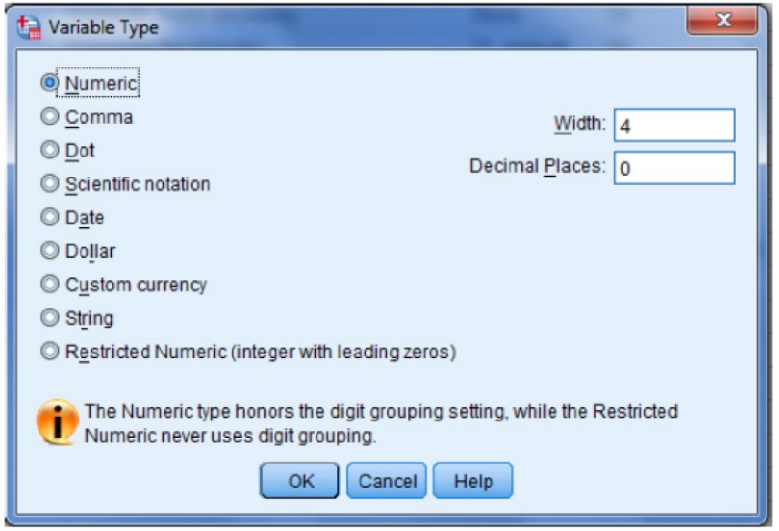
- Each variable (i.e. a question in a questionnaire) that we want to use for our research, in addition to its specific properties (i.e. possible answers - values, attributes), also carries a number of other important information that we must clearly characterize and describe in order to enter the analysis. This part is very important for the needs of the IBM SPSS software and filling it out allows this statistical software to display the results of the analysis in a format that other programs do not allow.
- By defining all the important characteristics of the individual variables, we create the so-called " **data matrix base** ", which is completely unique and specific to the IBM SPSS software. We will learn how to create it in practice in the third lesson, in the second lesson we will now explain what the individual characteristics mean and what forms they can take. These are eleven indicators that can be found in the columns of **the data window** of the " **Variable View" sheet** . These characteristics must be filled in for each variable that enters the analysis.

"Name" – variable name

- In this column, we assign each variable its name (title), but only in abbreviated form, according to its content focus and formal rules. Usually, the abbreviation of the variable that is to be analyzed by the given row is used, or the number of the question under which the given question in the questionnaire (or in another data collection tool) is hidden. However, certain rules must be followed for using abbreviations in the "Name" column, which are available in the box below.
- The name that we assigned to the variable in the "Name" column will subsequently be reflected in the data window of the "Data View" sheet as the name of the column that describes the variable.

"Type" – variable type

- In the "Type" column, we define the type of notation for each variable. We determine with what character / code / description we will write the individual answers to the given question in the "Data View" sheet. We do not formulate the type of notation manually, but through a dialog box that contains predefined options from which we can select the type of notation.
- The answer entry type dialog box is called up by clicking the mouse on the right end of the cell of the given column and row. After clicking, a dialog box will appear on the screen. Confirm the selection by clicking the "OK" button and then closing the dialog box with the cross in its upper right corner.



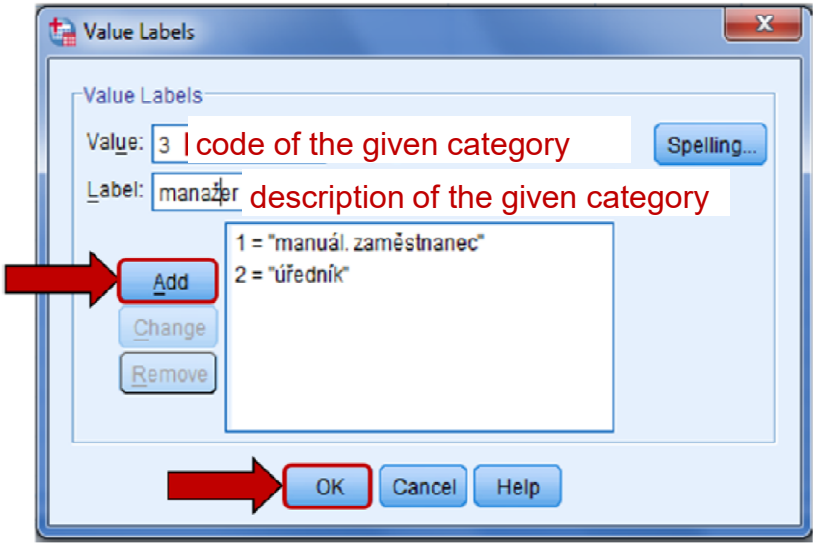
- When determining the type of variable, we choose between a numeric (digits and numbers) and a string (word) variant of the answer writing type. Based on this choice, the computer then evaluates during analysis whether the given variable can be used for numerical operations (in the case of the numeric variant) or not (in the case of the string variant).
- **Numerical variants include:**
  - Numeric: the most used option for most digits and numbers
  - Comma: the value of a thousand is recorded in the format 1,000.00
  - Dot: the value of a thousand is recorded in the format 1.000.00
  - Scientic notation: scientific notation
  - Dollar: numeric notation containing a dollar sign before the given value
  - Custom currency: you can enter the exchange rate of the desired currency
  - Restricted Numeric: numeric notation with a leading zero
- **The special numerical variant includes** (it is necessary to specify it further)
  - Date: it is also necessary to select the desired date format, which is offered in the window
- **The text variant includes:**
  - String: word description – answers are not encoded, but the entire word is described – used mainly when writing proper names

"Label" – variable description

- The “Label” column is used to specify the description of the variable name, which is listed in the “Name” column. The abbreviation of the variable in the “Name” column may not always be sufficiently concise or accurate, so we can better define it in the “Label” column.
- We don't have to follow any restrictions in this field, so we can assign better explanations to variable names, including spaces and diacritics.
- If we would like to write the same label in the "Label" field as the variable named in the "Name" field, we can, or we can leave the "Label" column for the given variable empty.
- So why is it appropriate to define a "Label"? The basic feature of the "Label" column is its ability to refine the printed results in the "Output". The labels of individual tables or graphs that we will create in the analyses take the names of the variables from the "Label" column, even though in the data window we work with the names from the "Name" column. If the "Label" field is empty, the outputs are described by the name from the "Name" field.

"Values" – description of the variable values

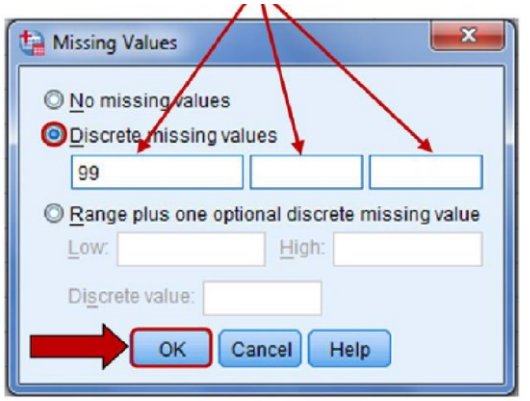
- The "Values" column is one of the most important columns for **categorized variables** . This is where we list the specific properties that the variable acquires and assign individual codes to them. The output tables and graphs will then not display the numerical codes under which we will enter the respondent's answers into the program, but they will be displayed in the verbal meaning, which will be encoded in the "Values" column in the "Variable View" sheet.
- To fill in the "Values" column, we must click the mouse on the right end of the cell, which will open a dialog box designed for writing individual answers and their codes. We then gradually fill in all the properties (answers, value, attribute) of the given variable into this table and assign them unique codes (two identical codes must not be repeated for two different answers). We will thus create a unique pair of code and verbal description, which we must then include in the list of other properties of a specific variable using the " **Add** " button. After entering the last possible answer and its code, we confirm with the "OK" button.



"Missing" – defining missing values

- In the "Missing" column, we define a placeholder code that replaces the missing answer in the questionnaire.
- If the respondent did not answer any of the questions (variables) (or answered implausibly, incorrectly), this field in the "Data View" data window cannot be left empty, but it is necessary to come up with a special category for these missing values in the individual properties (variants, values) of the given variable.
- We enter the code and verbal description of the missing answer in both the “Values” and “Missing” columns. We can define up to three types of missing answers here

it is possible to enter up to 3 types of missing responses



"Role" – role of the variable

- The "Role" column is loaded automatically in the basic IBM SPSS Statistics module and is practically not used in basic analyses.
- Its use mainly applies to working with the IBM SPSS Modeler extension model (a paid add-on to IBM SPSS Statistics), in which each variable is assigned a certain "role". The role then determines whether it is an input, target, or auxiliary variable, or whether analyses will be performed only with a certain part of the file or with its split.
- In the "Role" column menu we find the following options:
- a) “Input” – input variable
- b) **“Target”** – the target variable that we want to explain
- c) **“Both”** – the given variable fulfills both of the above roles (it is both an input and an explanatory variable)
- d) **"None"** – no role is assigned to this variable
- e) **"Partition"** – used only in special program models that work only with individual parts of the data in the file (the variable there defines which part of the file will be training, testing and validation)
- f) **“Split”** – used only for the extended version of IBM SPSS Modeler

Research progress

- determining the research topic,
- methodological preparation of the research (creation of the research objective, research questions, secondary research questions, research hypotheses, working hypotheses, operationalization process, which results in the creation of individual questions and answers in the data collection tool),
- implementation of preliminary research (i.e. verification of the data collection tool),
- adjusting the questionnaire according to the results of the preliminary research and
- implementation of a questionnaire survey.

Questionnaire data conversion

- After collecting the questionnaires from the respondents, it is necessary to number all the questionnaires and transcribe their answers into a pre-prepared data matrix in the IBM SPSS software. After entering all the answers into the software, it is necessary to check, clean and make basic adjustments to the data. After all these steps, you can proceed to the actual data analysis - according to the required procedures.
- In the overall context of creating a questionnaire, the basic rule applies - less is sometimes more, or the simpler the questions are, the easier it will be to analyze the answers.
- The creation of the data matrix (i.e. the transfer of questions from the questionnaire to the IBM SPSS software) is carried out in the data window, in the "Variable View" sheet, and must be the first, input step when creating a new data matrix. The data matrix (i.e. the transcription of respondents' answers to individual questions into the IBM SPSS software) is carried out in the data window, in the "Data View" sheet, which is predefined by the already created data matrix.

Creation of a basis for a data matrix according to a questionnaire template

- Converting data into electronic form is the most complex and time-consuming process of the entire statistical processing of collected information. This primary activity must be carried out very carefully, because if we make a mistake in creating the basis for the data matrix or in filling the matrix with data, the results we arrive at through the analyses will also be incorrect.
- It is always necessary to start in the data window, in the "Variable View" sheet, in which we gradually define the specific characteristics of individual variables (questions in the questionnaire) - thus creating the "basis for the data matrix".
- In the "Variable View" sheet, we always start with a row labeled "id". This is the first row of the data matrix, in which we will write the questionnaire number from which we will transcribe the collected data (it serves for retrospective control in case of incorrect data transcriptions).
- The next rows of the "Variable View" sheet already correspond to the order of questions from the questionnaire.
- We must pay the greatest attention to correctly filling in the "Values" field (this field is only filled in for nominal and ordinal variables) and correctly placing the variable in the measurement scale.
- If the questionnaire also contains more complex or completely ambiguous questions, we must pay close attention to the way they are entered into the data matrix.

Which questions may be problematic during transcription?

- Semi-open questions – within one question, a sub-question or a specification may also be posed.
- Complex questions – within one question, two or more content-wise different answers are required.
- Conditional questions – the answer to a given question is conditioned by the answer to a previous question (may not be completed by all respondents).
- Ranking questions – within one question, an answer based on assigning a set order to the values of individual responses is required.
- Batteries of questions – questions which make up a battery and are then used to create a new, loaded answer.
- Multiple-answer questions – for one question, a respondent may provide more than one

Semi-open questions

- Non-uniform questions are characterized by the fact that a sub-question or clarification is also asked within a single question. The respondent can therefore choose one of the offered answer options or specify their answer in the supplementary part of the question.
- In the process of transcription into electronic form in the IBM SPSS program, a question of this type needs to be divided into two parts - i.e. into the basic question, i.e. the main one (labeled e.g. o1) and its sub-question (labeled e.g. o1a), which will, however, be entered as a new separate variable on a new line (from one question we will have two questions - variables). However, the naming of the questions in the "Name" column is entirely up to the researcher who is preparing the data.
- However, when transcribing the supplementary part of the question (from our example of question o1a), we are not able to fill in the “Values” column in advance. Therefore, we fill in this open part of the semi-open question in the “Values” column only after collecting all the questionnaires from the respondents. The procedure is that we first go through all the paper questionnaires and write down on paper all the possible answers that the respondents wrote in the space for specifying the answer. We combine everything that can be combined (we combine answers with similar meanings - for example, language school and English school are the same for the purposes of this question) and code the individual answers. We only now write these codes, together with their explanation, in the “Values” column. This procedure must be followed for all open questions in the questionnaire!

Sources

DISMAN, M. (2014): How sociological knowledge is produced. Prague: Karolinum. ISBN 978-80-246-1966-8

REICHEL, J. (2009): Chapters in the methodology of social research. Prague: Grada Publishing as ISBN 978-80-247-3006-6.

KOUŘIL, P.; GABRHEL, V.; ŠIMEČEK, M.; SZABÓ, D.; TÖGEL, M. (2018): Construction of a sample set of traffic behavior surveys for urban planning purposes. *Urbanism and Territorial Development.* Volume XXI, 6/2018.