It would be helpful at the point to look at what an actual dataset is and how the information is displayed in Stata. In this section we will look at a dataset that I had previously prepared. I won't tell you anything about the dataset. Instead, we will load it into Stata and discover what sort of data it contains. You should download the file *dataset1* which is included with this lecture. Once you download it, double click on the file in windows in order open Stata with the data loaded in it. There are other ways to load data into Stata but at this point this is the simplest way and since the purpose of this section is to look at a dataset it doesn't make any sense to load the data in a more complicated way. Later on in the course we will see how we can load a dataset from inside Stata.

When you double click on the dataset, Stata will launch and the dataset will be automatically loaded. If you look at the area where the output is displayed, you will notice that Stata is telling you that it executed the **use** command. This command is used in order to tell Stata that you want to use a certain dataset. By double clicking on the dataset file, we actually told Stata that we want to use this specific dataset so Stata loaded it for us. You will also notice that the location of the file is written after the command use. This is because when we want to use a certain dataset we need to tell Stata where the dataset is stored. The text that you see will depend on where you had stored the file. My file was stored in the downloads folder so Stata has executed the command:

use "C:\Users\mozahemna.UNIVERSITY\Downloads\dataset1.dta"

Notice that we had to tell Stata the full path name for the file. This is because different files are stored in different folders and Stata needs to know in which exact folder the data file is stored. You don't have to worry about this at this point because the next section will explain the file system in Stata.

Going back to our output, we can also see the following line:

(This dataset contains the grades achieved by some students on certain courses)

This is actually the label which is applied to the data. If you remember, in the last lecture, I mentioned that it is good practice to label variables to make them more readable. The same is true about datasets. This dataset has been labelled and the label is telling us that the dataset contains information about student grades in some courses. So now we know what type of information is contained in the dataset that has been loaded into Stata.

If you look to the left-hand side at the rectangle that contains the commands that had already been executed, you can see the **use** command. If you want to execute it again, just double click on it. If you now look to the right-hand side at the area that displays the variables contained in the dataset, you can see that there are seven variables in this dataset, and they are named *id*, *gender*, *gpa*, *course_title*, *credits*, *grade*, and *semester*. Although the names are clear, a label is also included for each variable. For example, what is the difference between the variables *gpa* and *grade*? Looking at the labels of each variable, we can see that the variable *gpa* refers to the students' overall GPA in university, while the variable *grade* refers to the numeric grade achieved on a specific course.

If you click on the variable *id*, you see that the table on the bottom becomes populated with some information. The table shows that the variable id is of type *int* where int refers to the word integer.

An integer cannot contain a decimal point. This makes sense since student ids do not contain decimal points. If you next click on the variable *gpa*, you will notice that it is of type *double*. This type of variable can take on decimal values. Again this makes sense since a student can have a gpa of 78.33% (at least in some universities). If you next click on the variable *course_title*, you see that the variable is of type *str52* which is short for "a string that can contain up to 52 characters".

Next, click on the variable *gender*. This variable is a special case. Although you would expect gender to be a string variable, since a student can be a male or a female, you will notice that it is in fact of type *int*. You will also notice another thing and that is that the field "value label" in the table contains the text "gender2". At this point, I do not want to get into the details of this because later on we have a whole lecture dedicated to this special, but common, case.

Clicking on the variables in the graphical user interface in order to get information about each is one way to go about doing things. However, as I mentioned in the previous lecture, using commands in Stata is a better way of doing things. If we run the following command:

describe

Stata will produce output that gives us the information about the dataset and the variables in a visually attractive way. There is no need to click on each variable individually. We can see the label of the dataset, the variables names, and the variable labels. Stata also tells us that the dataset contains 5,594 observations or records.

What if we wanted to see the data as we do when we use Excel? This is very easy in Stata. In the toolbar on top there are two buttons that allow us to do that. The first button is labelled "Data Editor (Edit)" and the second button is labelled "Data Editor (Browse)". As the labels suggest, both buttons open the Data Browser, the first in editing mode, so the user can both view and edit the data, and the second in browsing mode, where the user can only see the data and no modifications can be made. If you click on the "Data Editor (Browse)" button, a new window will open. This window looks like what we see in an Excel file where the rows represent the records and the columns represent the variables. The names of the variables are written in the heading of each column. If you click in any cell and try to change its contents you will not be able to do so. This is because we have opened the data editor in browse mode. We can scroll through the records in order to get an idea of what the data looks like. We can also select records and copy them in order to paste them in Excel.

If you want to be able to modify the data, you can click on the "Edit mode" button on top. Stata will then ask you if you are sure that you want to do this. As you can see, Stata wants to make sure that you actually want to change the data and that you are not editing it by mistake. If you click "yes" Stata will allow you now to edit data. If you try to enter text in a variable that is of type *int* or *double*, Stata will not allow it.

Usually when you collect data you will not be able to collect data about every single variable in every single record. For example, you can see that there is no grade for the student with id 1 in the course which is titled "Business Communication Skills". In this case, the data is missing. Unlike Excel where missing data is represented by an empty cell, Stata visualizes an empty cell by having

a "." in it. Therefore, in Stata whenever you see a cell with a "." this would mean that the value for this variable for this particular record is missing.

This purpose of this lecture was to introduce you to what a dataset looks like in Stata. We saw that when we wanted to use a certain dataset that we had to tell Stata the location of the dataset. This idea of location is very important when you work with Stata, especially if you are working with multiple files where each file might reside in a different location. This is why the next lecture introduces you to the concept of directories and how to navigate through them in Stata.