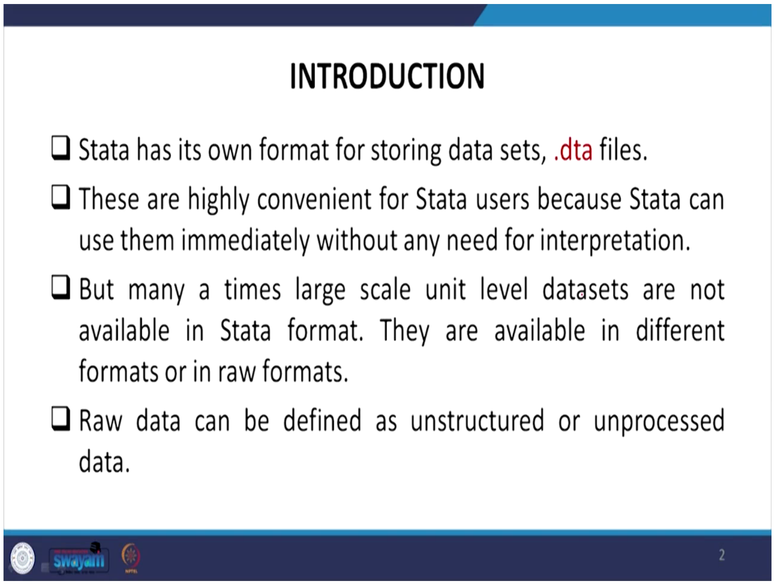
Handling Large-Scale Unit Level Data Using STATA
Professor. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture No. 21
Extraction in Stata-I

Friends once again I welcome you to my lecture on Handling Large-Scale Data Using Stata. This is as part of NPTEL MOOC module is sponsored by MHRD for our better use of data and better use of pedagogy in different set ups. Here we tried our best to enable you how to handle the unit level data with Stata and that too how to also interpret the data using Stata. This particular week is targeted for understanding some important intricacies of Stata handling of the unit level data. So, we have accordingly titled as handholding of unit level data.

(Refer slide Time: 01:18)



INTRODUCTION

- ❑ Stata has its own format for storing data sets, .dta files.
- ❑ These are highly convenient for Stata users because Stata can use them immediately without any need for interpretation.
- ❑ But many a times large scale unit level datasets are not available in Stata format. They are available in different formats or in raw formats.
- ❑ Raw data can be defined as unstructured or unprocessed data.

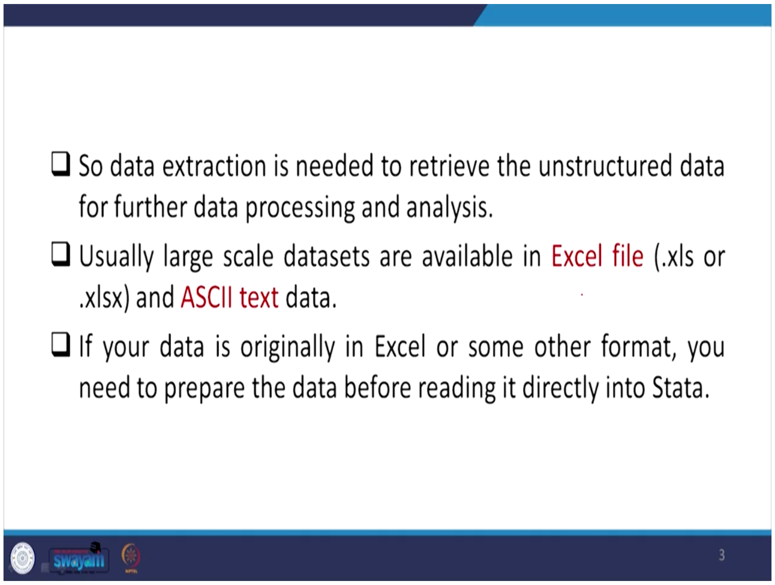
2

So, let me proceed for understanding the extraction as part of handholding. In these two lectures we are going to discuss the detailed information of data extraction and we will try our best today to stick to NSS or even economic census database and other data you can also explore. We will guide you in between for other databases wherever it is there.

So, let me just introduce you what do you mean by the extraction. Stata has its own format for storing datasets. We need to understand this background very clearly before extraction. Stata saves it in .dta files as I already guided you. These are highly convenient for Stata users, because

Stata can use them immediately without any need for interpretation. But many a times large scale unit level datasets are not available in Stata format. They are available in different formats or in, simply in raw formats. Raw data can be defined as generally an unstructured or unprocessed data.

(Refer Slide Time: 02:36)



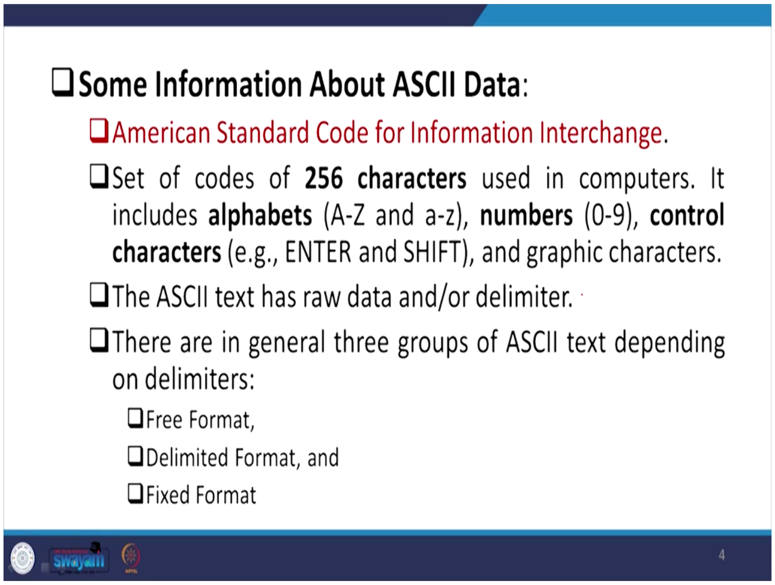
❑ So data extraction is needed to retrieve the unstructured data for further data processing and analysis.

❑ Usually large scale datasets are available in Excel file (.xls or .xlsx) and ASCII text data.

❑ If your data is originally in Excel or some other format, you need to prepare the data before reading it directly into Stata.

So, data extraction is needed to retrieve the unstructured data for further data processing and analysis. Usually large scale datasets are available in excel format or Excel file that is in .xls or xlsx and also in ASCII text data format. If your data is originally in Excel or other format, you need to prepare the data before reading it directly into Stata. So, you need to prepare it as per Stata requirement.

(Refer Slide Time: 03:05)



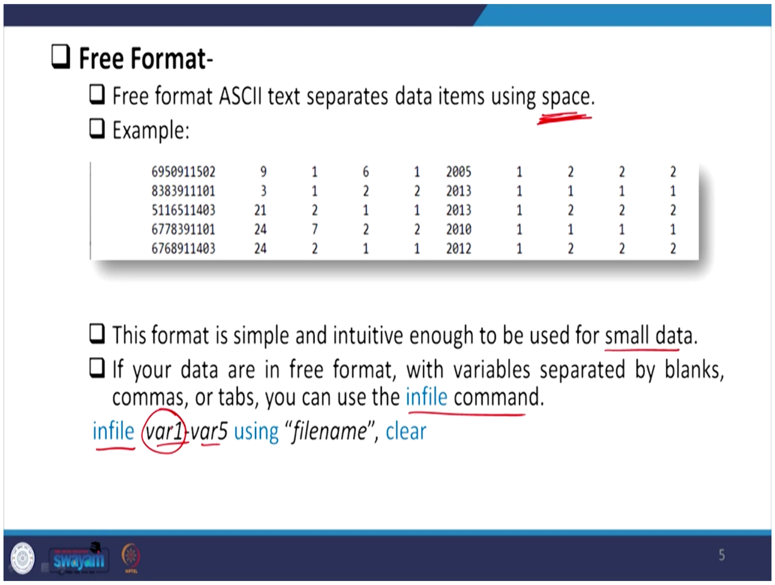
❑ **Some Information About ASCII Data:**

- ❑ American Standard Code for Information Interchange.
- ❑ Set of codes of **256 characters** used in computers. It includes **alphabets** (A-Z and a-z), **numbers** (0-9), **control characters** (e.g., ENTER and SHIFT), and graphic characters.
- ❑ The ASCII text has raw data and/or delimiter.
- ❑ There are in general three groups of ASCII text depending on delimiters:
  - ❑ Free Format,
  - ❑ Delimited Format, and
  - ❑ Fixed Format

So, what do you mean by ASCII data, it is called American Standard Code for Information Interchange. It is a set of codes of 256 characters used in computers. It includes alphabets that is A-Z in capital or in small, numbers from 0 to 9, control characters like enter and shift these are also there and also graphic characters. The ASCII text has raw data and/or delimiter. We are going to guide you on this in our next slide.

There are in general 3 groups of ASCII text depending on the delimiters which delimit the data or separates the data from one byte to another one. We are going to guide it clearly. So, these 3 are the following, free format, delimited format or fixed format. In this particular lecture we are going to guide you the handholding aspects of free format and delimited format. In the next lecture we are going to guide you the fixed format with the help of the core raw or the ASCII data available in India. What do you mean by free format data? Please do not miss any single point from our PPT. This is going to guide you very clearly.

(Refer Slide Time: 04:53)



❑ Free Format-

- ❑ Free format ASCII text separates data items using space.
- ❑ Example:

```
6950911502    9    1    6    1   2005    1    2    2    2
8383911101    3    1    2    2   2013    1    1    1    1
5116511403   21    2    1    1   2013    1    2    2    2
6778391101   24    7    2    2   2010    1    1    1    1
6768911403   24    2    1    1   2012    1    2    2    2
```

- ❑ This format is simple and intuitive enough to be used for small data.
- ❑ If your data are in free format, with variables separated by blanks, commas, or tabs, you can use the infile command.

infile *var1*-*var5* using "*filename*", clear

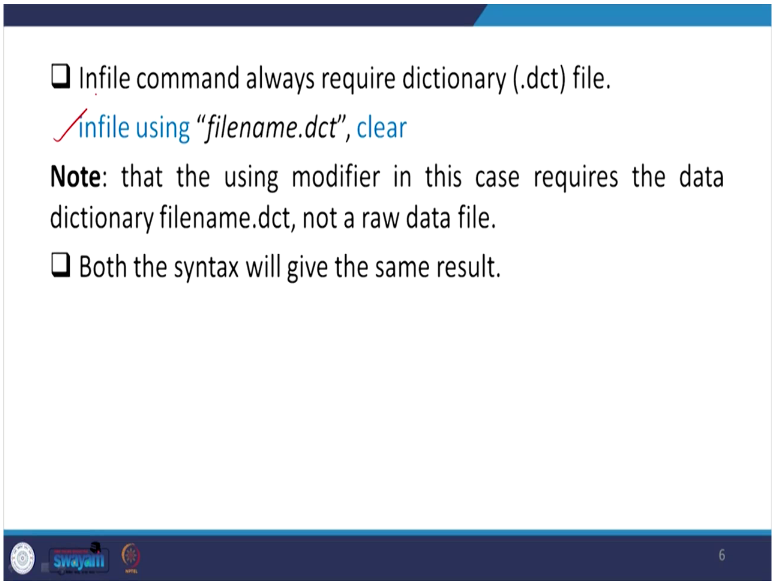
Free format ASCII text separates data items using a space. Note it very clearly that space is important in this case. Like in the example dataset which we have put it here for your understanding, the entry of data is delimited by space, isn't it? Here the space is different. There is different from one entry to another entry. So, this format is simple and intuitive enough to be used for small dataset.

So, if there is a small data, then generally they go for this, otherwise the other formats are important. If your data are in free format, which I just guided you, with variables separated by blanks or comma or tabs, you can use the command called infile, infile command will guide you to extract the data to your Stata format. Like if it is in free format and which you can understand by the space separation between the variables.

The command here is infile, this is very important, then you wanted to extract the variable, that there are so many variables out of that for my use, my research paper, I want variable 1 to 5 or maybe 1 to 10 whichever are important, you have to take the name of that particular variable, this variable, that variable, all 5 variables. For us I have written 5 variables, for you maybe 10 variables, you simply enter, name of those 10 variables.

As per given in that particular document, you cannot just name on your own. Whatever is given in the original document with that name, you enter that name here using then file name of that particular name which we are going to extract with a comma you need to enter, because comma will clear the earlier operation in that Stata if any is there.

(Refer Slide Time: 07:25)



❑ Infile command always require dictionary (.dct) file.

infile using "*filename.dct*", clear

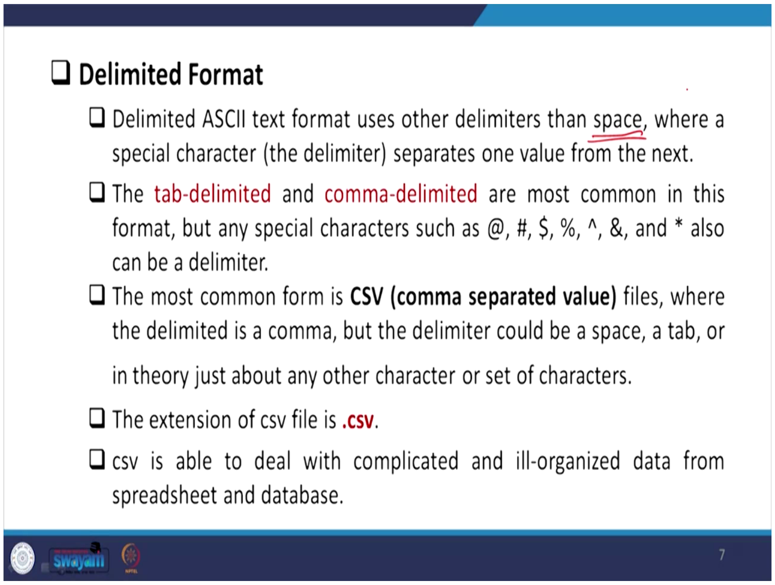
Note: that the using modifier in this case requires the data dictionary filename.dct, not a raw data file.

❑ Both the syntax will give the same result.

The infile command always require something that is important dot dct file, the dictionary file. If it is a free format, then infile requires the dot dct file for sure. File name with dot dct and infile entry is there and with this fixed space is there. When that space is there, then you apply this particular extraction command. One of the note here is that the using modifier in this case requires the data dictionary file that is in dot dct, not a raw data file. Both the syntax will give the same result.

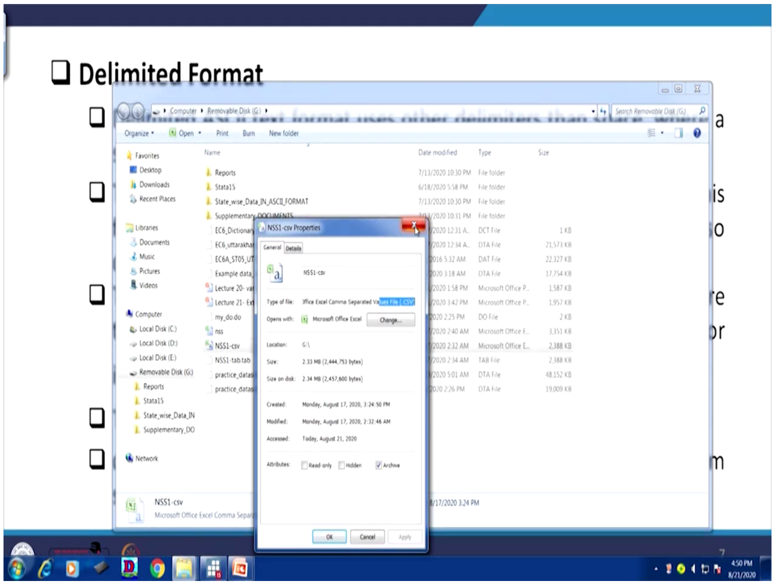
Like we have already mentioned here, you know the file name and you know the command name, that is in dot dct, this approach as well as the next page approach with the dictionary file that will give you the same result. I am not guiding you much because this type of data, NSS, even IHDS, even NFHS we are guiding is not available. So, we will guide you such type of data which we are going to use it for our purpose at large.

(Refer Slide Time: 08:36)



❑ Delimited Format

- ❑ Delimited ASCII text format uses other delimiters than space, where a special character (the delimiter) separates one value from the next.
- ❑ The tab-delimited and comma-delimited are most common in this format, but any special characters such as @, #, $, %, ^, &, and * also can be a delimiter.
- ❑ The most common form is **CSV (comma separated value)** files, where the delimited is a comma, but the delimiter could be a space, a tab, or in theory just about any other character or set of characters.
- ❑ The extension of csv file is **.csv**.
- ❑ csv is able to deal with complicated and ill-organized data from spreadsheet and database.



❑ Delimited Format

Computer ▸ Removable Disk (G:) ▸

Name	Date modified	Type	Size
Reports	7/13/2020 10:30 PM	File folder	
Stata15	6/18/2020 5:58 PM	File folder	
State_wise_Data_IN_ASCII_FORMAT	7/13/2020 10:30 PM	File folder	
Supplementary_DOCUMENTS	7/13/2020 10:31 PM	File folder	
EC6_Dictionar	/2020 12:31 A...	DCT File	1 KB
EC6_uttarakhar	/2020 12:34 A...	DTA File	21,573 KB
EC6A_ST05_U	2016 5:32 AM	DAT File	22,327 KB
Example data	2020 3:18 AM	DTA File	17,754 KB
Lecture 20- va	/2020 1:58 PM	Microsoft Office P...	1,587 KB
Lecture 21- Ex	/2020 3:42 PM	Microsoft Office P...	1,957 KB
my_do.do	2020 2:25 PM	DO File	2 KB
nss	/2020 2:40 AM	Microsoft Office E...	3,351 KB
NSS1-csv	/2020 2:32 AM	Microsoft Office E...	2,388 KB
NSS1-tab.tab	/2020 2:34 AM	TAB File	2,388 KB
practice_datas	/2020 5:01 AM	DTA File	48,152 KB
practice_datas	2020 2:26 PM	DTA File	19,009 KB

NSS1-csv Properties

NSS1-csv

Type of file: ffice Excel Comma Separated Values File (.CSV)

Opens with: Microsoft Office Excel

Location: G:\

Size: 2.33 MB (2,444,753 bytes)

Size on disk: 2.34 MB (2,457,600 bytes)

Created: Monday, August 17, 2020, 3:24:50 PM

Modified: Monday, August 17, 2020, 2:32:46 AM

Accessed: Today, August 21, 2020

Attributes: □ Read-only □ Hidden ☑ Archive

NSS1-csv Microsoft Office Excel Comma Separ 8/17/2020 3:24 PM

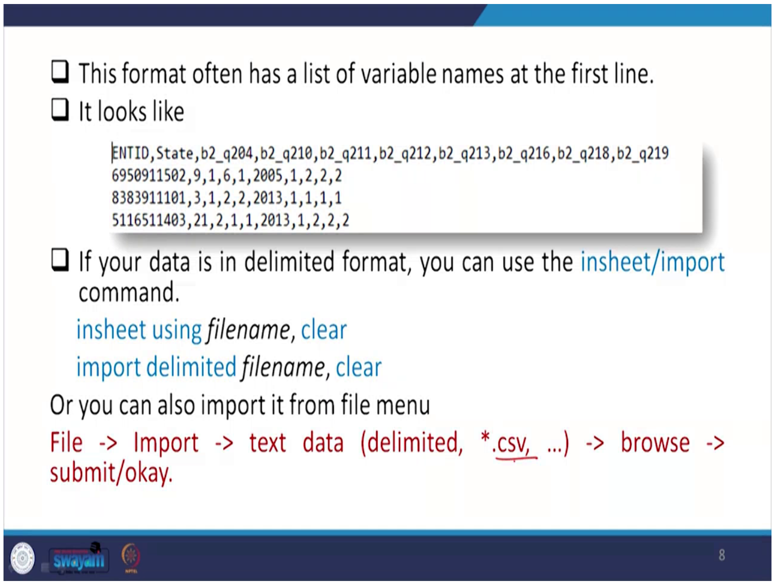
So, let us come to delimited format. This is even better, a better format than compared to the previous one. We are going to tell you the advantages and disadvantages of each of the format. First one is free format, we told you. Delimited ASCII text format uses other delimiters than space that is very important to note. So, earlier case it is only space, but other delimiters are important. What are the delimiters here? The tab-delimited so you should have a tab-delimited file also or comma-delimiter. So, comma-delimiter, I will tell you what kind of tab-delimited and comma-delimited we are going to guide you.

So, comma-delimited are most common. These two are most common format in this set up, but any special character such as: at, hash, dollar, power and or asterisk also can be a delimiter, but the most common are tab. Just simply tab then another variable or comma then another variable, no need to have space. The most common form is CSV file that is also called comma separated file, usually it is in comma format, where the delimiter is a comma, but the delimiter could be a space, a tab or in theory just about any other character or set of characters, any other. So, we are broadly going to talk about CSV or tab-delimited approach.

The extension of CSV file is written like dot csv. I think in our data I will guide you. This is our data. look at this CSV file. if I just right click here, we need to right click on this, so properties, so it will show you here at the end, it gives you also some information, otherwise, .csv information generally comes here .csv. Here you get the dot csv information. Otherwise, it also written if you enter the extension file here in your size, type, data and it will also show you in one of the column. If you take all those options, it also shows you the CSV option. There are many approaches to understand whether it is in CSV or in any other one.

Let me proceed. So, what I will do, I will guide you step by step. CSV is able to deal with complicated and ill-organized data from spreadsheet and data sheet. Generally, spreadsheet cannot organize everything. Therefore, this data compress as well as organize. Save with limited space and give you a better format and for analysis. So, all the big, large scale data adopts these techniques.

(Refer Slide Time: 11:53)



- ❑ This format often has a list of variable names at the first line.
- ❑ It looks like

```
ENTID,State,b2_q204,b2_q210,b2_q211,b2_q212,b2_q213,b2_q216,b2_q218,b2_q219
6950911502,9,1,6,1,2005,1,2,2,2
8383911101,3,1,2,2,2013,1,1,1,1
5116511403,21,2,1,1,2013,1,2,2,2
```

- ❑ If your data is in delimited format, you can use the insheet/import command.

  insheet using *filename*, clear

  import delimited *filename*, clear

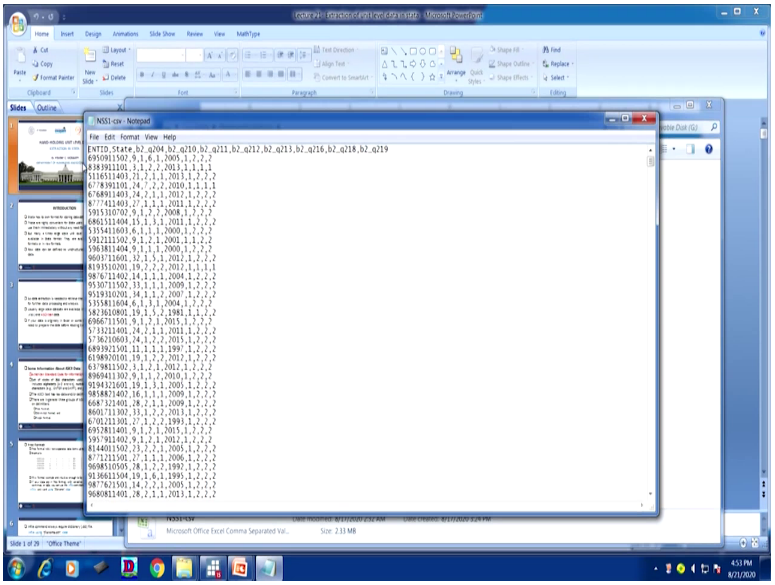
Or you can also import it from file menu

File -> Import -> text data (delimited, *.csv, ...) -> browse -> submit/okay.

8

This format often has a list of variable names at the first line. Most importantly you need to understand that this, it looks like this. First line the first row generally contains the variable names. Here in our data we will show you the enterprise ID, state, other variable, these are variable names from the beginning. Then other rows are the observations or the cases.

(Refer Slide Time: 12:24)

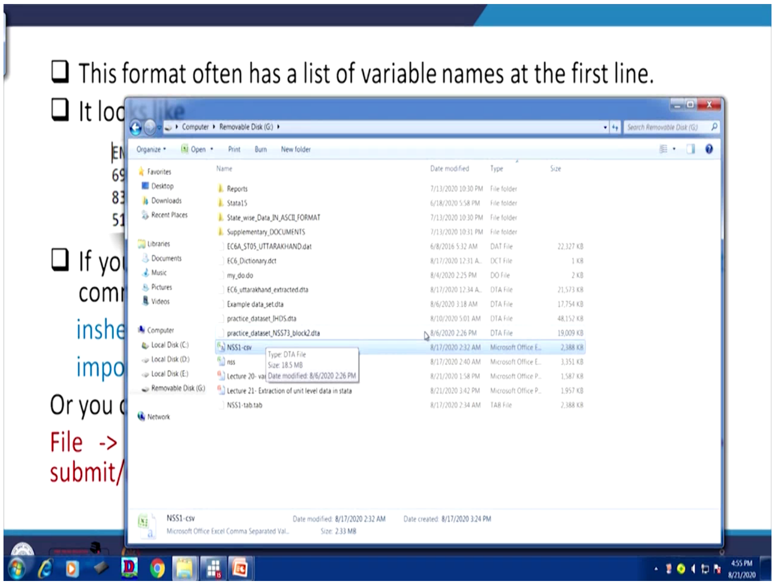


```
ENTID,State,b2_q204,b2_q210,b2_q211,b2_q212,b2_q213,b2_q216,b2_q218,b2_q219
6950911502,9,1,6,1,2005,1,2,2,2
8383911101,3,1,2,2,2013,1,1,1,1
5116511403,21,2,1,1,2013,1,2,2,2
6778391101,24,7,2,2,2010,1,1,1,1
6768911403,24,2,1,1,2012,1,2,2,2
8777411403,27,1,1,1,2011,1,2,2,2
5915310702,9,1,2,2,2008,1,2,2,2
6861511404,15,1,3,1,2011,1,2,2,2
5355411603,6,1,1,1,2000,1,2,2,2
5912111502,9,1,2,1,2001,1,1,2,2
5963811404,9,1,1,1,2000,1,2,2,2
9603711601,32,1,5,1,2012,1,2,2,2
8193510201,19,2,2,2,2012,1,1,1,1
9876711402,14,1,1,1,2004,1,2,2,2
9530711502,33,1,1,1,2009,1,2,2,2
9519310201,34,1,1,2,2007,1,2,2,2
5355811604,6,1,3,1,2004,1,2,2,2
5823610801,19,1,5,2,1981,1,1,2,2
6966711501,9,1,2,1,2015,1,2,2,2
5733211401,24,2,1,1,2011,1,2,2,2
5736210603,24,1,2,2,2015,1,2,2,2
6893921501,11,1,1,1,1997,1,2,2,2
6198920101,19,1,2,2,2012,1,2,2,2
6379811502,3,1,2,1,2012,1,2,2,2
8969411302,9,1,1,2,2010,1,2,2,2
9194321601,19,1,3,1,2005,1,2,2,2
9858821402,16,1,1,1,2009,1,2,2,2
6687321401,28,2,1,1,2009,1,2,2,2
8601711302,33,1,2,2,2013,1,2,2,2
6701211301,27,1,2,2,1993,1,2,2,2
6952811401,9,1,2,1,2015,1,2,2,2
5957911402,9,1,2,1,2012,1,2,2,2
8144011502,23,2,2,1,2005,1,2,2,2
8771211501,27,1,1,1,2006,1,2,2,2
9698510505,28,1,2,2,1992,1,2,2,2
9136611504,19,1,6,1,1995,1,2,2,2
9877621501,14,2,2,1,2005,1,2,2,2
9680811401,28,2,1,1,2013,1,2,2,2
```

I can show you that, so it is here. I will go to our file we can open it, the CSV file. This is the CSV file. But we have opened in Stata, we need to open it, check it with right click, open with at the top, open with notepad. You look at the data which the sample is given to you, the example is given to you. It is entID, state, then all the variables from the beginning and all other variables are separated by comma. You look at very carefully all are separated by comma. all the observations we have already shortlisted are given here. All the observations are given for our use.

Let me proceed further. So, what I have mentioned here to you that we have already discussed you this. So, this is the one I have already clarified. If your data is in delimited format, you can use the insheet command or import command. So, the command is given here. Insheet using file name then you can able to extract. I will show you right now. Insheet using that file name and then that will extract the data. Otherwise, import delimited file and clear. If you do that, you can able to extract it. Let me just have a check from our data.

(Refer Slide Time: 14:20)



❑ This format often has a list of variable names at the first line.

❑ It loo

EN
69
83
51

❑ If yo
comr
inshe
impo

Or you c
File ->
submit/

Computer ▸ Removable Disk (G:) ▸

Name	Date modified	Type	Size
Reports	7/13/2020 10:30 PM	File folder	
Stata15	6/18/2020 5:58 PM	File folder	
State_wise_Data_IN_ASCII_FORMAT	7/13/2020 10:30 PM	File folder	
Supplementary_DOCUMENTS	7/13/2020 10:31 PM	File folder	
EC6A_ST05_UTTARAKHAND.dat	6/8/2016 5:32 AM	DAT File	22,327 KB
EC6_Dictionary.dct	8/17/2020 12:31 A...	DCT File	1 KB
my_do.do	8/4/2020 2:25 PM	DO File	2 KB
EC6_uttarakhand_extracted.dta	8/17/2020 12:34 A...	DTA File	21,573 KB
Example data_set.dta	8/6/2020 3:18 AM	DTA File	17,754 KB
practice_dataset_IHDS.dta	8/10/2020 5:01 AM	DTA File	48,152 KB
practice_dataset_NSS73_block2.dta	8/6/2020 2:26 PM	DTA File	19,009 KB
NSS1-csv	8/17/2020 2:32 AM	Microsoft Office E...	2,388 KB
nss	8/17/2020 2:40 AM	Microsoft Office E...	3,351 KB
Lecture 20- va	8/21/2020 1:58 PM	Microsoft Office P...	1,587 KB
Lecture 21- Extraction of unit level data in stata	8/21/2020 3:42 PM	Microsoft Office P...	1,957 KB
NSS1-tab.tab	8/17/2020 2:34 AM	TAB File	2,388 KB

Type: DTA File
Size: 18.5 MB
Date modified: 8/6/2020 2:26 PM

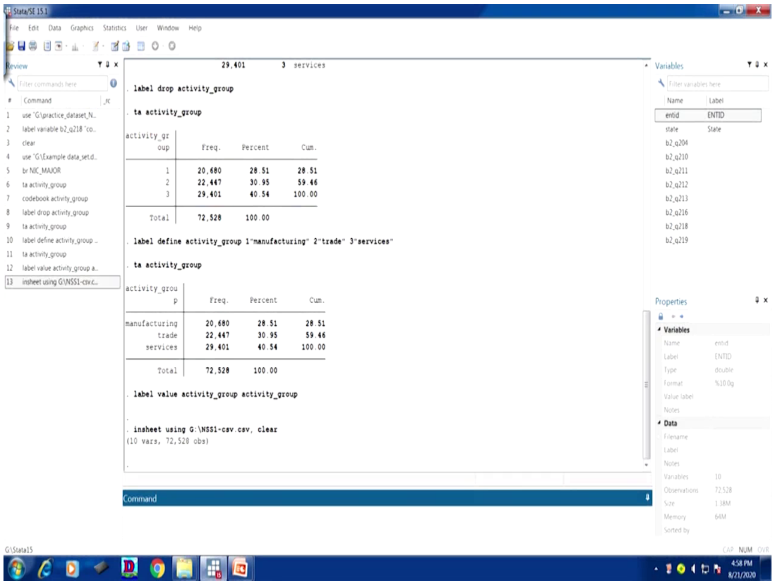
NSS1-csv Date modified: 8/17/2020 2:32 AM Date created: 8/17/2020 3:24 PM
Microsoft Office Excel Comma Separated Val... Size: 2.33 MB

The data is in, so we need to open in Stata, but now what I do I will open the Stata for you. We have to go by the command. This command if you enter, but you require, most importantly your data must be having the file name with csv. If the data is in csv then you can able to extract it. But if the data is not in csv or data look at, the original data is not in csv, it is in that file, the original data.

In fix command like this you can open. So, like this one is our data. If I just, the csv it is there, if I use that extension you can able to command. we have converted it as per our use. Originally the data is not, if you just go to the NSS or even economy sensors, you would not get the data in CSV format. You have to convert it with by file, save as in CSV you can able to convert it to CSV format if you wanted to test whether it is getting extracted or not. So, we have converted for our use.

So, our data here, now you can see it is, I have already shown you just couple of minutes back that it is NSS1 hyphen csv we have named it. I need the path of that. This is in G file; I have to use that for my command for a better extraction. This is, what I will do, I will copy this or I will type it insheet using then file name this I have already shown you. Now what I will do, I will, where is that command, it is here.

(Refer Slide Time: 16:24)



```
#   Command
1   use "G:\practice_dataset_N...
2   label variable b2_q218 "co...
3   clear
4   use "G:\Example data_set.d...
5   br NIC_MAJOR
6   ta activity_group
7   codebook activity_group
8   label drop activity_group
9   ta activity_group
10  label define activity_group ...
11  ta activity_group
12  label value activity_group a...
13  insheet using G:\NSS1-csv.c...
```

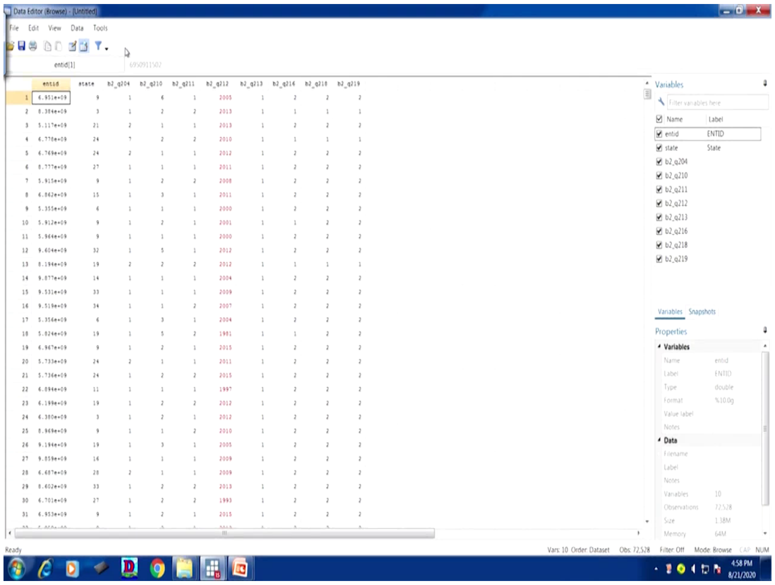
```
                 29,401          3  services

. label drop activity_group

. ta activity_group

activity_gr |
        oup |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     20,680       28.51       28.51
          2 |     22,447       30.95       59.46
          3 |     29,401       40.54      100.00
------------+-----------------------------------
      Total |     72,528      100.00

. label define activity_group 1"manufacturing" 2"trade" 3"services"

. ta activity_group

activity_grou |
            p |      Freq.     Percent        Cum.
--------------+-----------------------------------
manufacturing |     20,680       28.51       28.51
        trade |     22,447       30.95       59.46
     services |     29,401       40.54      100.00
--------------+-----------------------------------
        Total |     72,528      100.00

. label value activity_group activity_group

.
. insheet using G:\NSS1-csv.csv, clear
(10 vars, 72,528 obs)

.
```

Name	Label
entid	ENTID
state	State
b2_q204	
b2_q210	
b2_q211	
b2_q212	
b2_q213	
b2_q216	
b2_q218	
b2_q219	

So, this is insheet, you just check very carefully that insheet, you have to mention very clearly, do not miss any single word, using then file name. So, how to get the file name, file name is, I have already shown you G, it is in G path. So, it is in, it is here, so it is in G path. So, let me copy that first or what you do, just click here or right click and copy it. You just take that to your Stata. You using file name right click and paste it. So, this is your path has come. What else you require, you require the file name also.

So, what is the file name, file name is here, just copy that file name. So, right click and copy. So, I have already copied. So, what I do, I will try to paste it here. So, paste. The extension should be given. So, since it is in dot csv, so csv should be there. There are some original data already given, so you need to enter comma and clear. If you do not clear then there will be some overlapping issues, there will be some problem comma and clear. Type clear, if you click it clear then enter. You will get the information. Now your data is ready. This has been extracted.

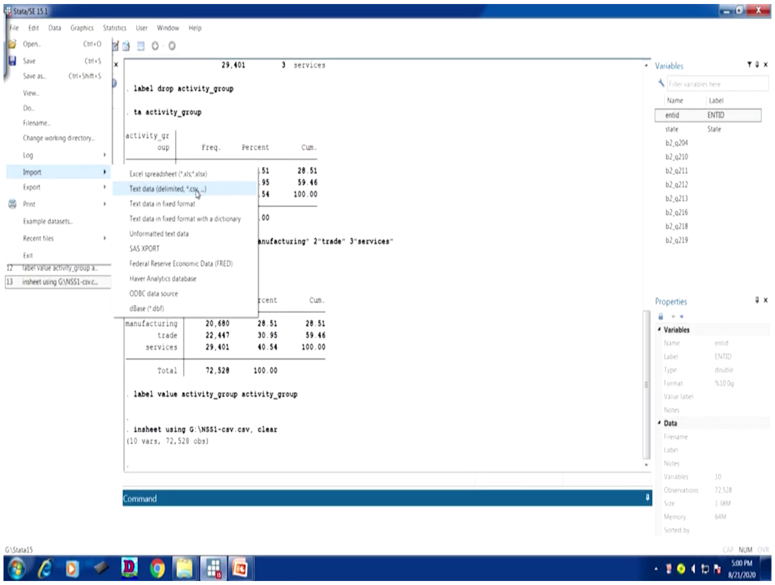
(Refer Slide Time: 18:20)



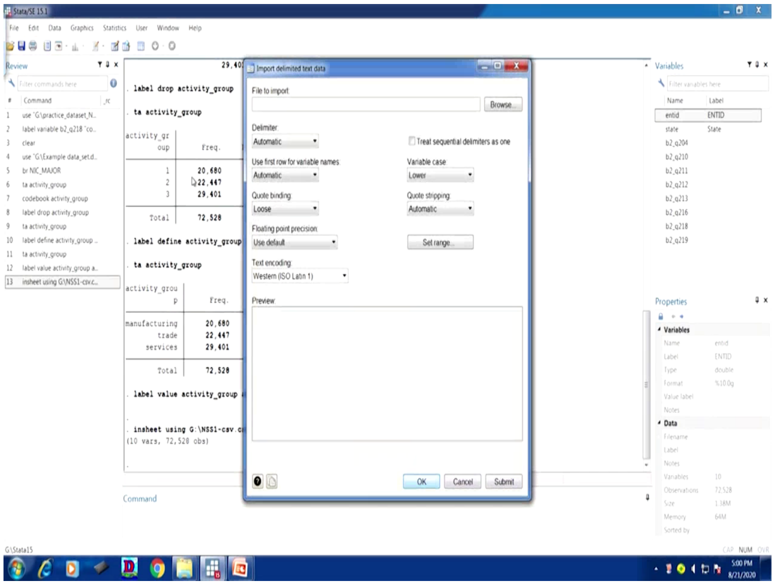
Data Editor (Browse) - [Untitled]

entid[1] 6950911502

	entid	state	b2_q204	b2_q210	b2_q211	b2_q212	b2_q213	b2_q216	b2_q218	b2_q219
1	6.951e+09	9	1	6	1	2005	1	2	2	2
2	8.384e+09	3	1	2	2	2013	1	1	1	1
3	5.117e+09	21	2	1	1	2013	1	2	2	2
4	6.778e+09	24	7	2	2	2010	1	1	1	1
5	6.769e+09	24	2	1	1	2012	1	2	2	2
6	8.777e+09	27	1	1	1	2011	1	2	2	2
7	5.915e+09	9	1	2	2	2008	1	2	2	2
8	6.862e+09	15	1	3	1	2011	1	2	2	2
9	5.355e+09	6	1	1	1	2000	1	2	2	2
10	5.912e+09	9	1	2	1	2001	1	1	2	2
11	5.964e+09	9	1	1	1	2000	1	2	2	2
12	9.604e+09	32	1	5	1	2012	1	2	2	2
13	8.194e+09	19	2	2	2	2012	1	1	1	1
14	9.877e+09	14	1	1	1	2004	1	2	2	2
15	9.531e+09	33	1	1	1	2009	1	2	2	2
16	9.519e+09	34	1	1	2	2007	1	2	2	2
17	5.356e+09	6	1	3	1	2004	1	2	2	2
18	5.824e+09	19	1	5	2	1981	1	1	2	2
19	6.967e+09	9	1	2	1	2015	1	2	2	2
20	5.733e+09	24	2	1	1	2011	1	2	2	2
21	5.736e+09	24	1	2	2	2015	1	2	2	2
22	6.894e+09	11	1	1	1	1997	1	2	2	2
23	6.199e+09	19	1	2	2	2012	1	2	2	2
24	6.380e+09	3	1	2	1	2012	1	2	2	2
25	8.969e+09	9	1	1	2	2010	1	2	2	2
26	9.194e+09	19	1	3	1	2005	1	2	2	2
27	9.859e+09	16	1	1	1	2009	1	2	2	2
28	6.687e+09	28	2	1	1	2009	1	2	2	2
29	8.602e+09	33	1	2	2	2013	1	2	2	2
30	6.701e+09	27	1	2	2	1993	1	2	2	2
31	6.953e+09	9	1	2	1	2015	1	2	2	2



```
                    29,401          3  services

. label drop activity_group

. ta activity_group

activity_gr |
        oup |      Freq.     Percent        Cum.
------------+-----------------------------------
            |          .51       28.51
            |          .95       59.46
            |          .54      100.00
------------+-----------------------------------
            |          .00

nufacturing" 2"trade" 3"services"

                   rcent        Cum.
------------+-----------------------------------
manufacturing |     20,680       28.51       28.51
      trade |     22,447       30.95       59.46
   services |     29,401       40.54      100.00
------------+-----------------------------------
      Total |     72,528      100.00

. label value activity_group activity_group

.

. insheet using G:\NSS1-csv.csv, clear
(10 vars, 72,528 obs)

.
```



#	Command	_rc
1	use "G:\practice_dataset_N...	
2	label variable b2_q218 "co...	
3	clear	
4	use "G:\Example data_set.d...	
5	br NIC_MAJOR	
6	ta activity_group	
7	codebook activity_group	
8	label drop activity_group	
9	ta activity_group	
10	label define activity_group ...	
11	ta activity_group	
12	label value activity_group a...	
13	insheet using G:\NSS1-csv.c...	

```
                          29,40

. label drop activity_group

. ta activity_group

activity_gr
        oup |      Freq.
------------+-----------
          1 |     20,680
          2 |     22,447
          3 |     29,401
------------+-----------
      Total |     72,528

. label define activity_group

. ta activity_group

activity_grou
            p |      Freq.
--------------+-----------
manufacturing |     20,680
        trade |     22,447
     services |     29,401
--------------+-----------
        Total |     72,528

. label value activity_group 

.
. insheet using G:\NSS1-csv.c
(10 vars, 72,528 obs)

.
```

Import delimited text data

File to import: [Browse...]

Delimiter: Automatic

□ Treat sequential delimiters as one

Use first row for variable names: Automatic

Variable case: Lower

Quote binding: Loose

Quote stripping: Automatic

Floating point precision: Use default

[Set range...]

Text encoding: Western (ISO Latin 1)

Preview:

[OK] [Cancel] [Submit]

Name	Label
entid	ENTID
state	State
b2_q204	
b2_q210	
b2_q211	
b2_q212	
b2_q213	
b2_q216	
b2_q218	
b2_q219	

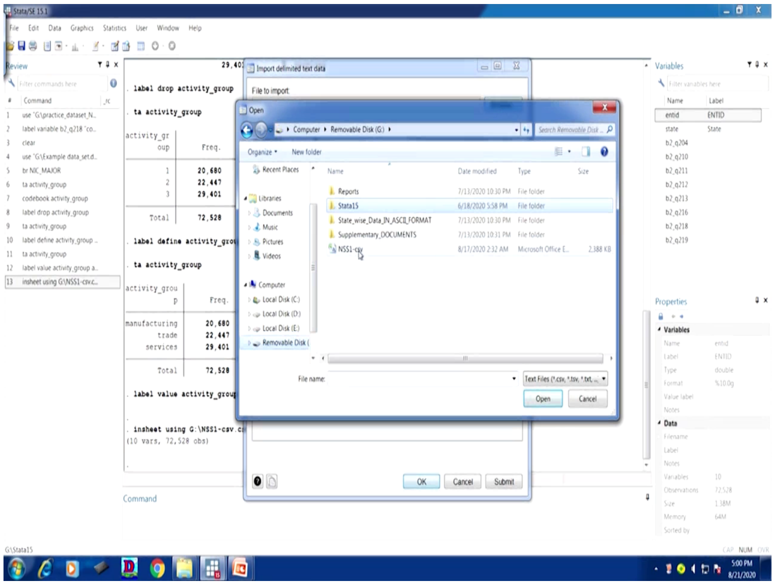
Properties

Variables

Name	entid
Label	ENTID
Type	double
Format	%10.0g
Value label	
Notes	

Data

Filename	
Label	
Notes	
Variables	10
Observations	72,528
Size	1.38M
Memory	64M
Sorted by	

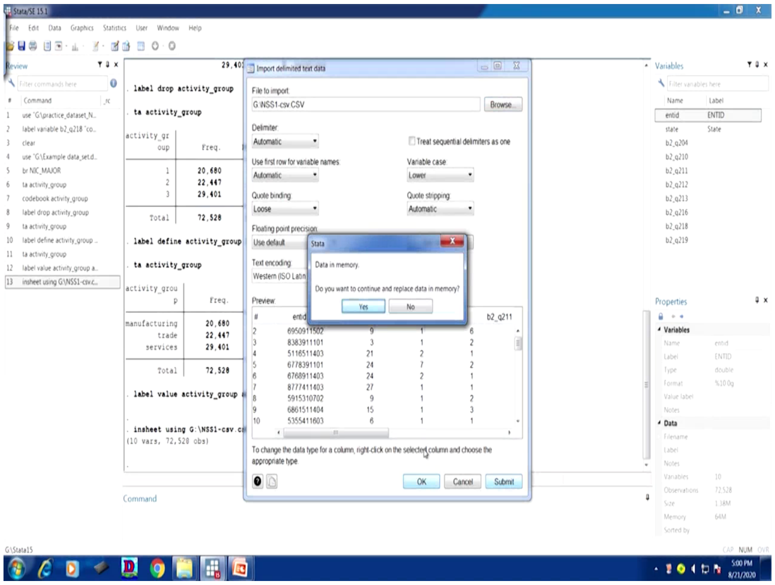


Name	Date modified	Type	Size
Reports	7/13/2020 10:30 PM	File folder	
Stata15	6/18/2020 5:58 PM	File folder	
State_wise_Data_IN_ASCII_FORMAT	7/13/2020 10:30 PM	File folder	
Supplementary_DOCUMENTS	7/13/2020 10:31 PM	File folder	
NSS1-csv	8/17/2020 2:32 AM	Microsoft Office E...	2,388 KB

File name: | Text Files (*.csv, *.tsv, *.txt, ...)

[Open] [Cancel]

[OK] [Cancel] [Submit]



```
1   use "G:\practice_dataset_N...
2   label variable b2_q218 "co...
3   clear
4   use "G:\Example data_set.d...
5   br NIC_MAJOR
6   ta activity_group
7   codebook activity_group
8   label drop activity_group
9   ta activity_group
10  label define activity_group ...
11  ta activity_group
12  label value activity_group a...
13  insheet using G:\NSS1-csv.c...
```

```
                       29,401

. label drop activity_group

. ta activity_group

activity_gr
        oup |      Freq.
------------+-----------
          1 |     20,680
          2 |     22,447
          3 |     29,401
------------+-----------
      Total |     72,528

. label define activity_group

. ta activity_group

activity_grou
            p |      Freq.
--------------+-----------
manufacturing |     20,680
        trade |     22,447
     services |     29,401
--------------+-----------
        Total |     72,528

. label value activity_group

.
. insheet using G:\NSS1-csv.c
(10 vars, 72,528 obs)

.
```

Import delimited text data

- File to import: G:\NSS1-csv.CSV
- Delimiter: Automatic
- □ Treat sequential delimiters as one
- Use first row for variable names: Automatic
- Variable case: Lower
- Quote binding: Loose
- Quote stripping: Automatic
- Floating point precision: Use default
- Text encoding: Western (ISO Latin

Preview:

#	entid			b2_q211
2	6950911502	9	1	6
3	8383911101	3	1	2
4	5116511403	21	2	1
5	6778391101	24	7	2
6	6768911403	24	2	1
7	8777411403	27	1	1
8	5915310702	9	1	2
9	6861511404	15	1	3
10	5355411603	6	1	1

To change the data type for a column, right-click on the selected column and choose the appropriate type

Stata

Data in memory.

Do you want to continue and replace data in memory?

So, we can see your data also. This has been extracted. This is a sample for you. This is insheet command we have shown you. Similarly, you can also apply another command that is import delimited file name then clear. The same approach, the same result you are going to get and you can able to extract it.

Another approach is you require either csv data or the delimited data. So, in our file look at this. So, we have, this is our G file. We have the NSS delimited data that is in. So, what I do, I will click and show you how to get. So, you have to go first to the file then import then text it, then you have to carefully observe CSV, if it is in CSV format, then you click on CSV, then submit okay you will get it like this. Then file import then here this is the one text data CSV format.

You click on this. Browse is there. You simply click on browse. You have to go to the data. This is the CSV we have shown you. submit. you need to replace it. Replace command we have not given it. So, do you want to continue and replace the data, yes. You need to replace the previous one then it will replace and you can get it, where the replace one. Since the data we already extracted and it is already available here. Otherwise, you just simply start a fresh document, start the fresh Stata page and go by this file and import then CSV data. It will extract it as per your use.

Now we are going to discuss the detailed information about the other format that is fixed format. So, far we guided you the free format and the delimited format and how you can able to do it, but as I told you, subject to the availability of data. If the data is in CSV or delimited file then you need to apply this command that is insheet or import delimited. Any command you do it, it will extract correctly the way we have shown it. Our next lecture will be on fixed format data and its extraction using the core dataset. With this, let me close here. Thank you.