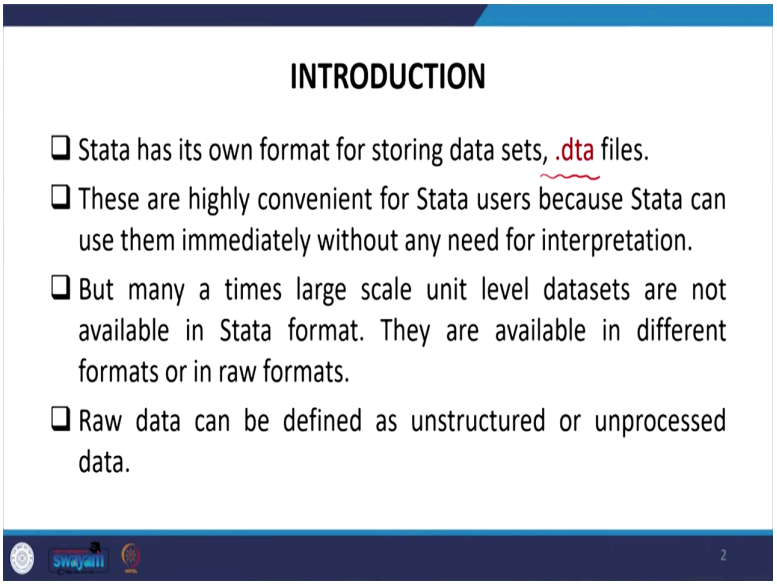
Exploring Survey Data on Health Care
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

Lecture - 22
Data Extraction from ASCII Format

Welcome friends once again to the NPTEL MOOC module on Handling Health Care Data. We are on the fifth week and trying to understand the survey data and their analysis. On this lecture we will be giving you the hand holding of NSS data and their most important aspect is on data extraction especially from the ASCII format database. Myself, Dr. Pratap C. Mohanty, I have been teaching research methodology and these type of courses over 6-7 years.

So, I will be most happy to address all your queries. So, let us move forward and understand what you mean by this data set and how we can set the tuning for understanding it for final use. Starting from the very beginning of understanding data especially in Stata. Stata has it is own format for storing data sets in *.dta* format that has to be noted very clearly. So, *.dta* if that extension you see in the file; that means, this is the Stata data.

(Refer Slide Time: 01:40)

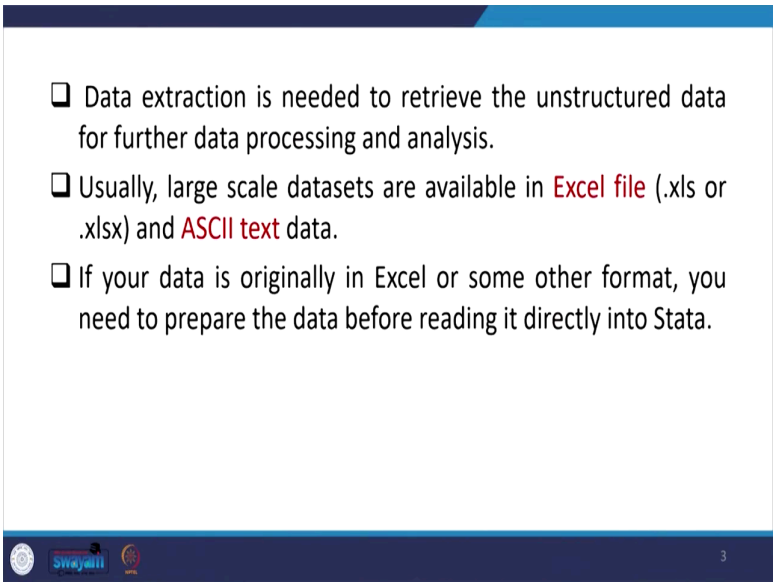


These are highly convenient for Stata users because Stata can use them immediately without any need for interpretation. But many a times large scale unit level data sets are not available

in Stata format. They are available in different formats or in raw formats or in ASCII data format or in simply in notepad format.

Raw data can be defined as unstructured or unprocessed data. So, lots of understanding is required and based on that we can go for processing the data before final analysis. Data extraction is in fact, needed to retrieve the unstructured data for further data processing and analysis.

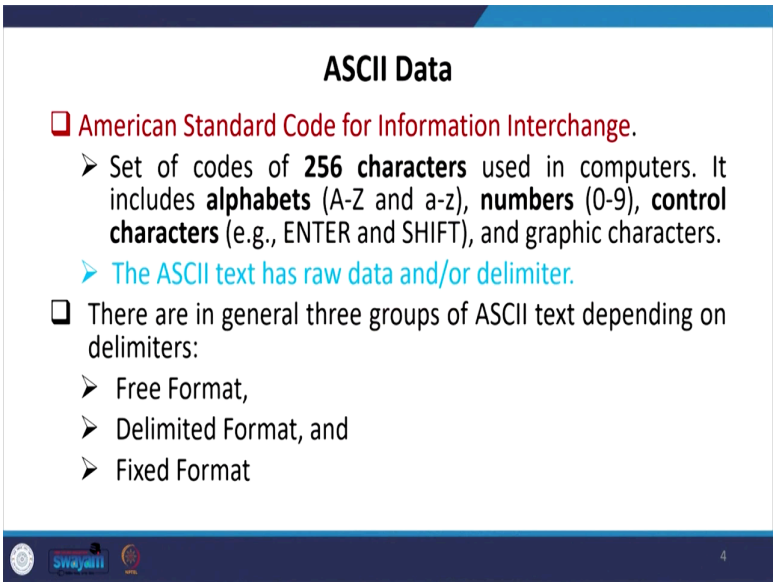
(Refer Slide Time: 02:27)



Usually, the large-scale survey data sets are available in Excel file, either in .xls format or in .xlsx format or in ASCII text data format. If your data is originally in Excel or some other format, you need to prepare the data before reading it directly into Stata. Because it is labeling and other thing that must be checked first and then you can import it to Stata.

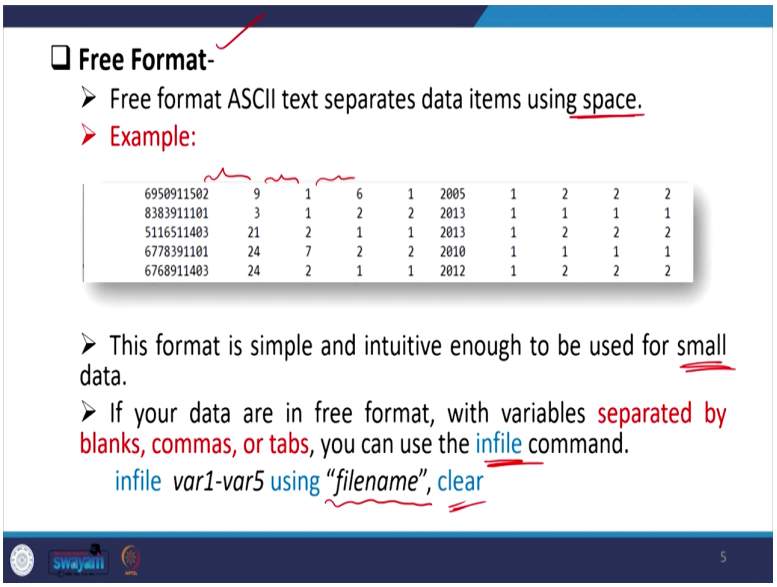
So, then what do you mean by ASCII text data? Some clarification we are trying to give here. These are basically a kind of American Standard Code for Information Interchange that is the full form. It has set of codes usually of 256 characters saved in computers. These includes alphabets, numbers then control characters like enter, shift, space then and some graphic characters as well.

(Refer Slide Time: 03:36)



The ASCII text datasets are raw data, or they are also called delimiters. There are in general three groups of ASCII text depending upon delimiters. Three different types; first one is called free format, then delimited format and fixed format. So, these three we are just going to explain you one by one for your clarity.

(Refer Slide Time: 04:07)



First, in this particular format, that is, free format the data considers spaces in between. There are clear designated spaces involved to separate the variable or the information. So, in the

example dataset you can see the spaces are so clearly identified and these differentiate the entries.

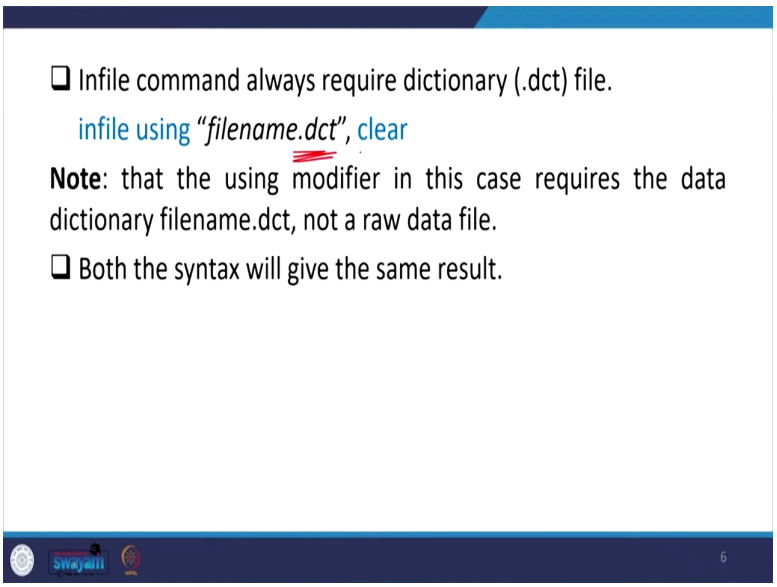
And the format is simple and intuitive enough to be used for small data, usually this is applicable for small data entries. If your data are in free format with variables separated by blanks or comma or tabs you can use one command for extraction that is called *infile*. *Infile* command is useful for this kind of free format data.

So, in that case you can use infile then variable and you can specify the variable names, variable 1 or till 5 or it is byte position if it is given. Then, 'where is your source of data?' should be clearly given. Because if any earlier stored data is available, it may not able to extract it.

At this time since we are not going to show you with *infile* format data or free format data. We are not going to show the exact operation but just for your clarity we have mentioned the command and we will operate the infixed format data and on that basis you can also understand this kind of data set if any, you are having in future.

So, *infile* command always requires a dictionary file through which you can able to extract the data.

(Refer Slide Time: 06:11)



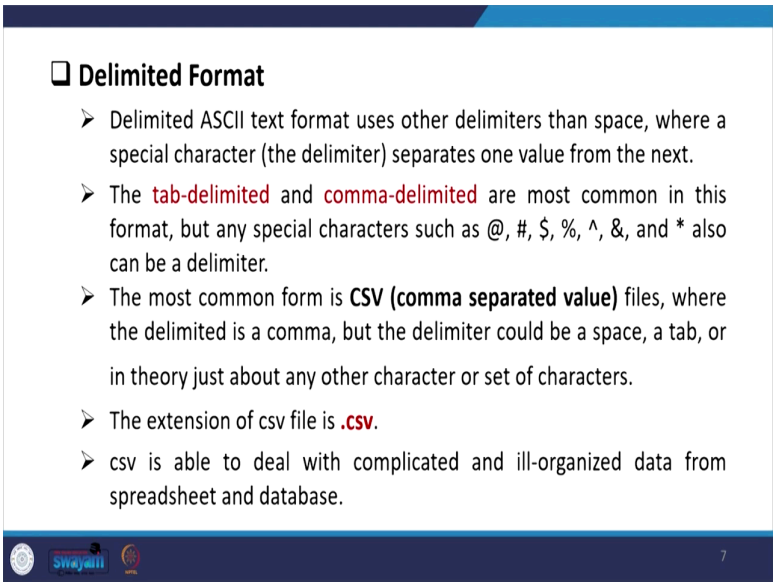
We are going to define dictionary file shortly: What do you mean by dictionary file?, How dictionary file is defined?, How a dictionary should be made that Stata is going to read before

it is extraction? Then we need to note here, that the using modifier in this case requires the data dictionary file that is *.dct* not a raw data file.

So, once you have opened the Stata and your dct file has already specified the source of the data, it is path name has already been discussed in the dct file or in the dictionary file. Then after the using command your dct file is most important, no need to further open your raw data. Both the syntax will give the same result.

On the first page i.e., before this page, we have given the variable name and their byte position. Without the dictionary file on this page we are guiding you through the dictionary file both will give you the same result.

(Refer Slide Time: 07:19)



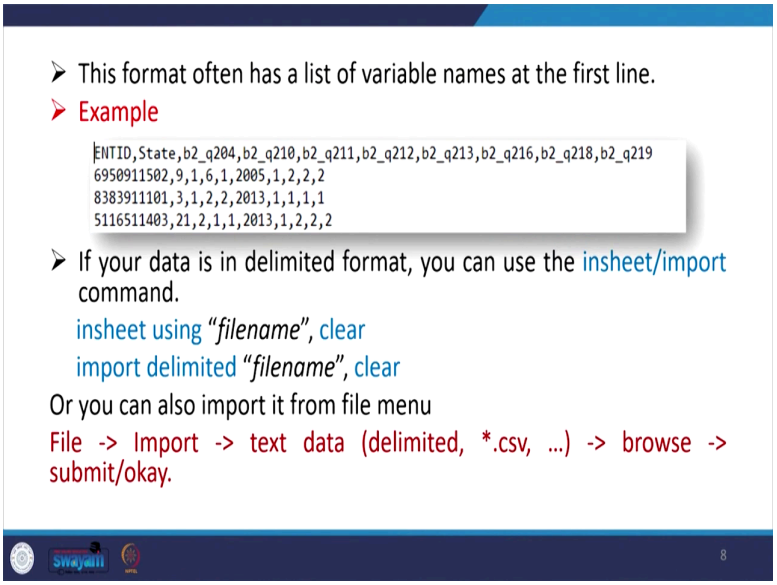
Next format is called delimited format. Delimited format is also important and these are also one form of text ASCII data format. Delimited ASCII text format uses other delimiters than space or comma. There are other delimiters could be used. So, like tab delimited, comma delimited two are most commonly used.

And besides that, some special character can be used as delimiters. Dollar symbol could be also given at the end of the variable or the beginning of the variable to separate the variable name. So, the delimiters could be of these: you may be asked in the exam that, 'can these are considered to be delimiters in some of the data in your objective question?'.

So, be careful about it. You have to be very clear that these are going to be useful when we have a delimited format data. The most common form of delimited data is called CSV that is that 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 characteristic or set of characters used for separating it that is why, it is called comma separated delimiters.

The extension of csv file is called *.csv*. *.csv* is able to deal with complicated and ill-organized data from spread sheet and database.

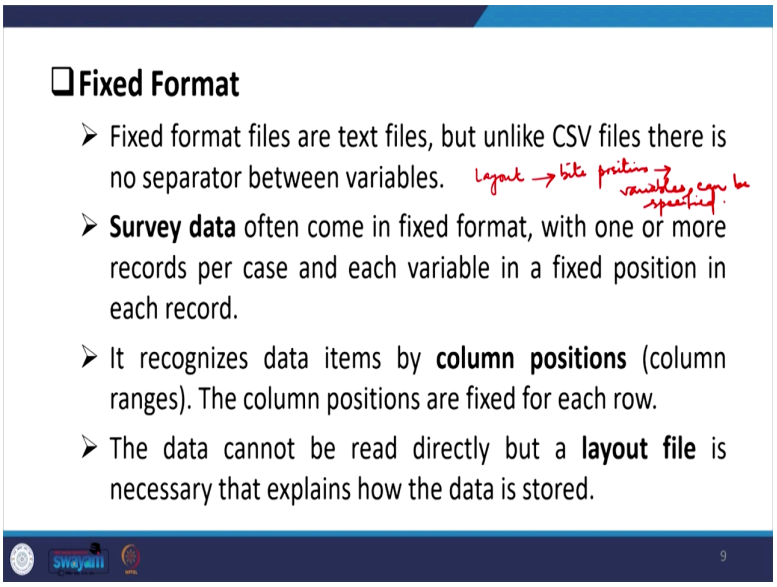
(Refer Slide Time: 09:12)



The sample is given here like here a list of variables are given and these are separated by comma. And so, delimiter here is our comma but in case of our first example that is in this free format data we discuss about space is the delimiter. If your data is in delimited format you can use the *insheet* command or import command. Both the commands are useful; *insheet* or *import* command will give you the same result. But if you are using import then you have to discuss the delimited import delimited file that has to be mentioned.

If it is *insheet* then *insheet* using "file name" has to be mentioned. Otherwise in Stata, if you go to file then you click on the very first important primary buttons you will find file. And then on the drop down menu you click on import, then text data, then delimited or *.csv* format, then browse and submit, you can get the result.

(Refer Slide Time: 10:39)



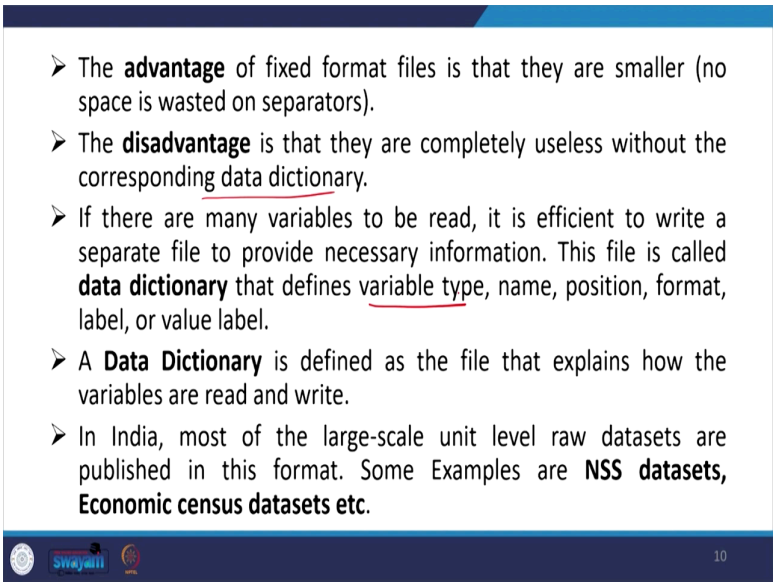
Fixed format, the third one which is mostly used and we are also going to use it now for practical purposes as well for showing you the direct extraction result. So, fixed format files are usually of text files, but unlike csv file there is no separator between variables. So, all the data points could be conjoined could be set together without any space or any delimiters.

So, those delimiters even if it is not there, still we can read between the entries because of their space identification in the layout file. We have shown you earlier about layout file. Layout file gives very systematic presentation of your byte positions, byte positions then, even if there is no comma separated files still, you can able to specify the variable names.

Then survey data often come in these formats with one or more records per case and each variable in a fixed position in each record. This recognizes data items by column position i.e., which column position these are entered and also their column ranges. The column positions are fixed for each row. The data for every row files you will find the same position that is why we are saying it is fixed.

The data cannot be read directly, but a layout file is required. So, layout file we have mentioned here. So, we have shown it in our data when I discussed on the very first week of our module.

(Refer Slide Time: 12:56)



The advantage of fixed format file is that they are smaller and no space is wasted or no separator is required. Therefore, the byte space it consumes because of that file is usually smaller. The disadvantage of this is that they are completely useless without the corresponding data dictionary.

If data dictionary is not defined correctly then it is actually very difficult. Since we have large number of variable and there are entries requires, a layout file specifies the byte position by its column distance or column position.

So, every time it is very difficult to remember and operate. So, a data dictionary is required. If corresponding data dictionary is not supported then this is going to be very cumbersome. If there are many variables to be read it is efficient to write a separate file to provide necessary information, the file is actually called data dictionary.

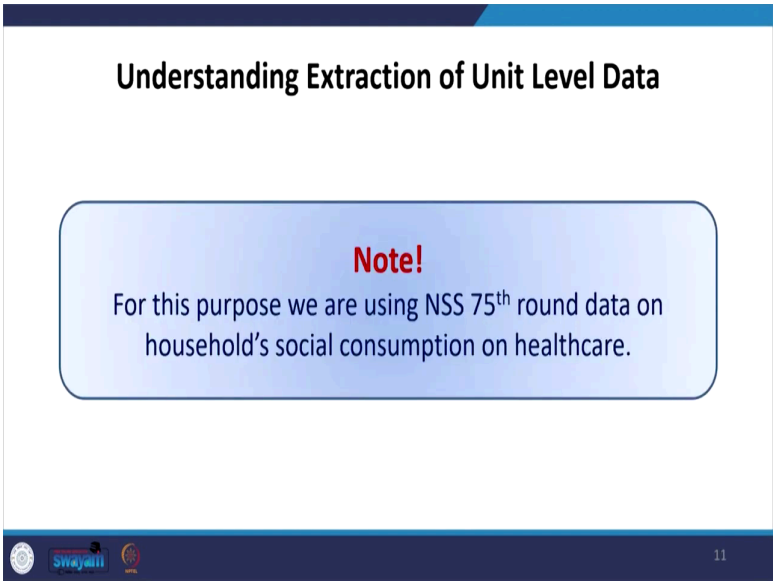
So, this is what I was saying. If so many variables are mentioned by their byte position it is better to prepare a separate file with all those details. We are going to guide it shortly. Those two type of file are called data dictionary. This defines variable type, name, position, format, level, value level etc.

So, data dictionary includes first, the variable type, whether these are in string or numeric. Then name of the variable we wanted to specify in the dictionary file. Then its position, byte

position and its duration then we can also specify their format or also we can specify their value level etc. can also be given on the dictionary file as well.

A data dictionary is defined as the file that explains how the variables are read and write. In India, most of the large scale unit level raw data sets are published in this format some examples are NSS data sets, then economic census data sets etc.

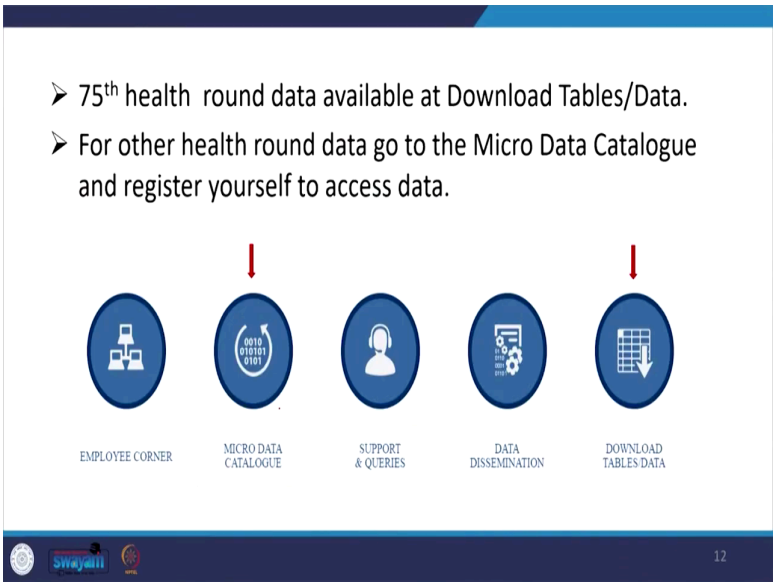
(Refer Slide Time: 15:19)



Understanding extraction of unit level data; we have used NSS 75th round for extraction which we have said from the beginning of our module that we will be exercising and this module on NSS 75th round.

Similarly, you can also extend it to any other round. This round is on health care and as on NSS 75th round there are in fact, two sub-round that is, one is on health care another is on education.

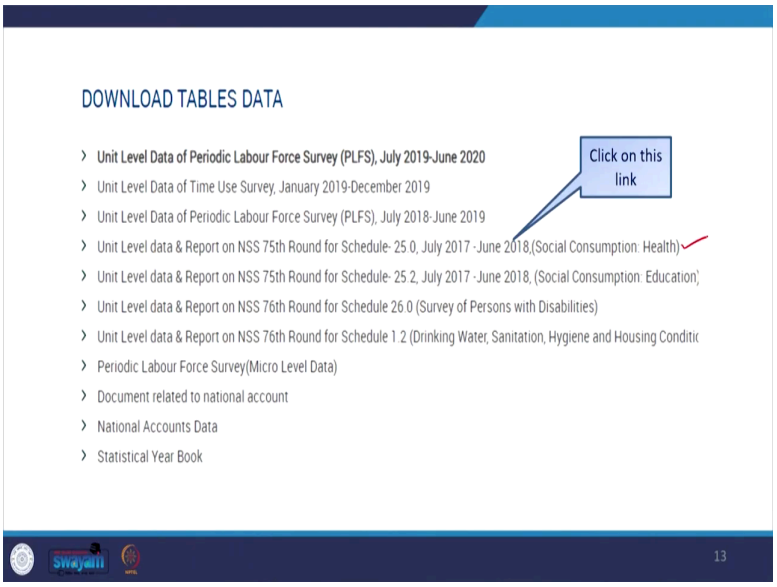
(Refer Slide Time: 16:12)



So, micro data you need to get for the raw data. Once you click on this, the micro data catalogue gives you all available micro data of NSS, but once you click on this download tables data you will be redirected to the concerned page. But as we have already said it requires a minimum i.e., a very basic registration which will be further approved based on your registration.

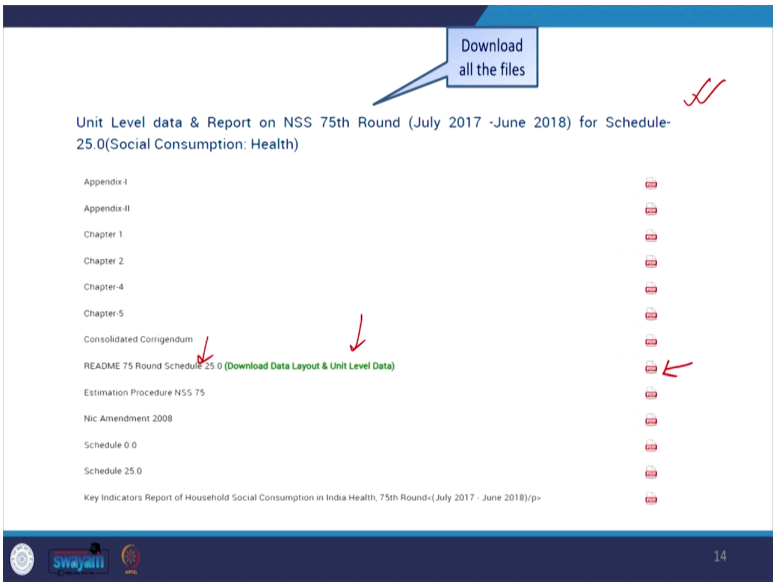
Usually it is not paid, it is freely available for the public. For other health round data, go to the micro data catalogue and register yourself to access these data. With the page which we have already guided earlier will look like this. At this moment we have also other data sets available like PLFS.

(Refer Slide Time: 17:12)



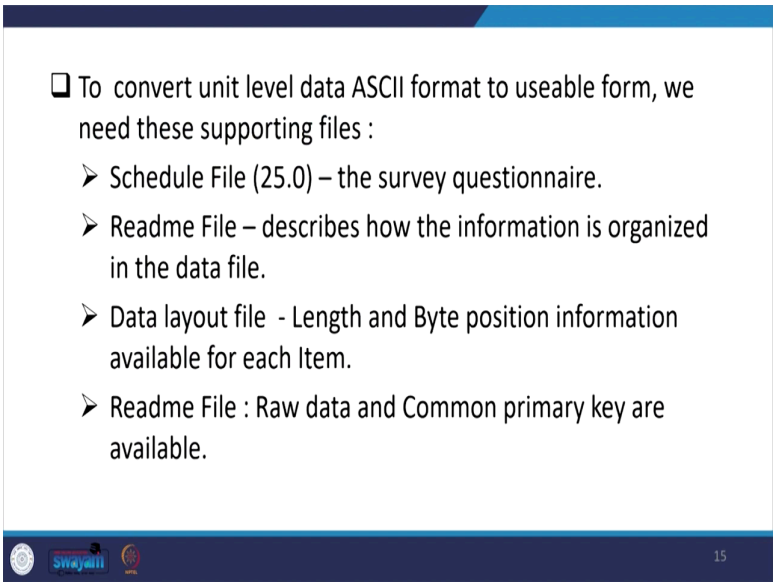
We are here trying to download this one. So, you just click on this link, and you will be downloading the required data.

(Refer Slide Time: 17:22)

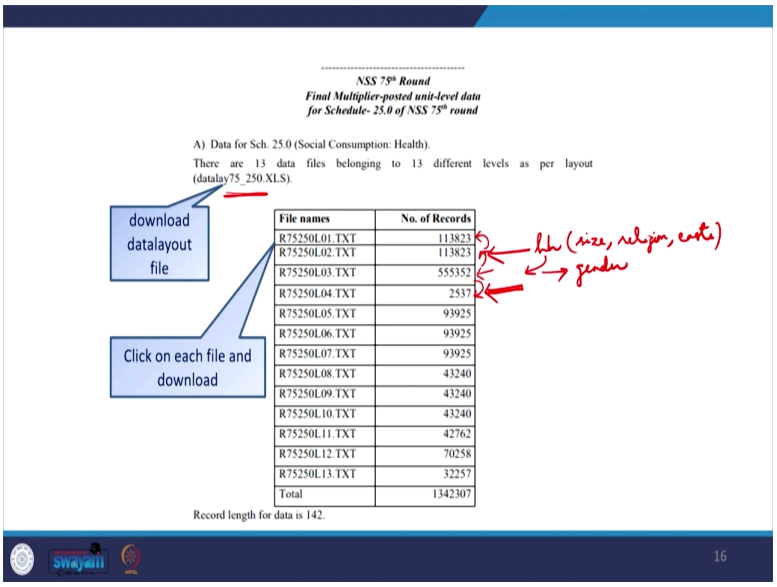


Now, next one is to open, once you click on that you will be redirected to this particular page. You will open this page and within this you have a link called layout and README file. So, you just click on this, a PDF will be opened in another page. On this PDF you can see a different guidance.

(Refer Slide Time: 17:50)



(Refer Slide Time: 17:51)



So, once you click on this you will be shown a complete page. I will also show it here. So, you simply click on each of the file, on this NSS 75th data there are 13 levels information. Why levels are important? I have already told you that the ministry has categorized the different subsets of information.

Like household is separated from individual, then within individual they will also categorize for those persons who accessed health care or having some forms of element over last 15 days or in last 365 days or there is certain information about deaths.

So, these information are important. Just for your clarity once again; once you see these data having number of cases is 555352. And this is in fact, the highest, that does indicate that this must be an individual file where it includes all the individuals.

Then before that usually NSS provides on the second this one is your household information. Before that the very basic information about the recoding of the different record its FSU units, Second stage stratum units etc.

Then now you might be confused that why it is so less? It is because very less number of responses must have been there since this is on deaths. Information about deaths i.e., those who have died in last 365 days and if the family member reported it. So, accordingly the data reported is very less.

So, a quick check for you that once you open this README file you should download these as well. Keep your cursor over here on data layout, you will be redirected to download. Similarly, on each of these 13-txt file you will have to download all those details.

So, once again I am guiding those who have already done it earlier like to convert the unit level data ASCII format to useable form, for that we need these supporting files as well. So, supporting files are very-very essential. So, like schedule file is required. Schedule means what? I have already guided in one of the lectures is on schedule versus questionnaire. Schedule is a systematic and complete document of information.

It is not just the questionnaire; it is also including a structured format of questionnaire. The Readme file which I have just shown you and a data layout file which you can able to download from the link which gives the byte position of all the information or the all the variables. Then the Readme file basically gives raw data.

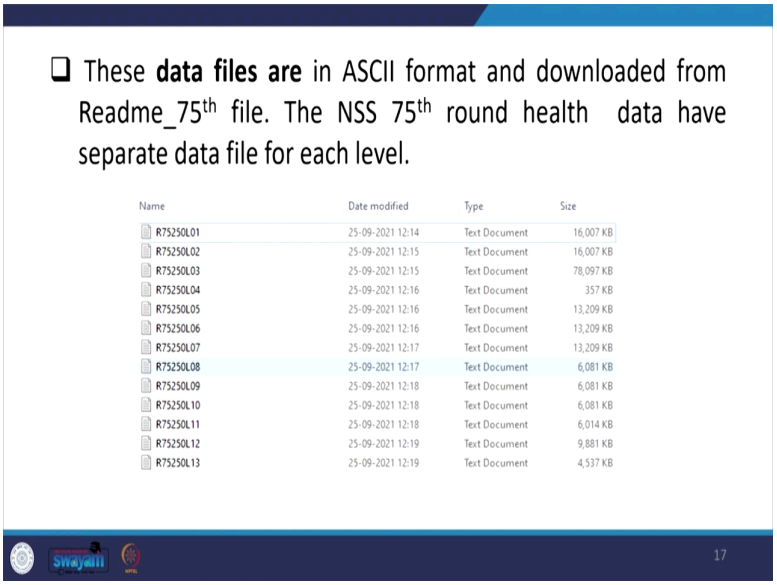
So, we need to download the raw data very clearly and you need to mark the common primary key. Common primary keys are essential because these holds in merging different blocks. So, different blocks of information why it is required because if you have simply here for example, 113823 are the household.

If you do not merge with the individual file, you cannot get the how complete information. Like from the household you will get household size, household religion and household caste.

Then you have to mix it with or joint or merge the data of individual information like individual gender. Then there are so many individual identification like individual education.

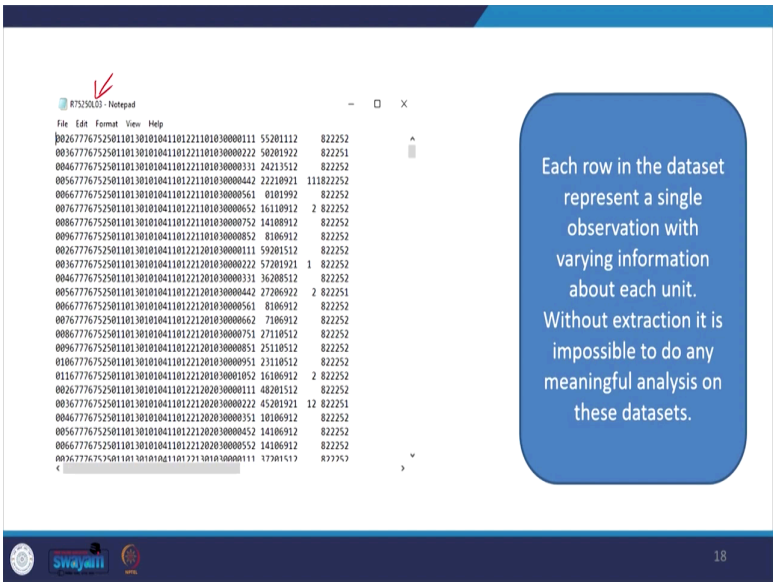
From the household you can get the standard of living or the income, NSS usually give a expenditure information. But in different sub blocks also or blocks of information you will also get information about their expenditure on different heads as well.

(Refer Slide Time: 22:22)



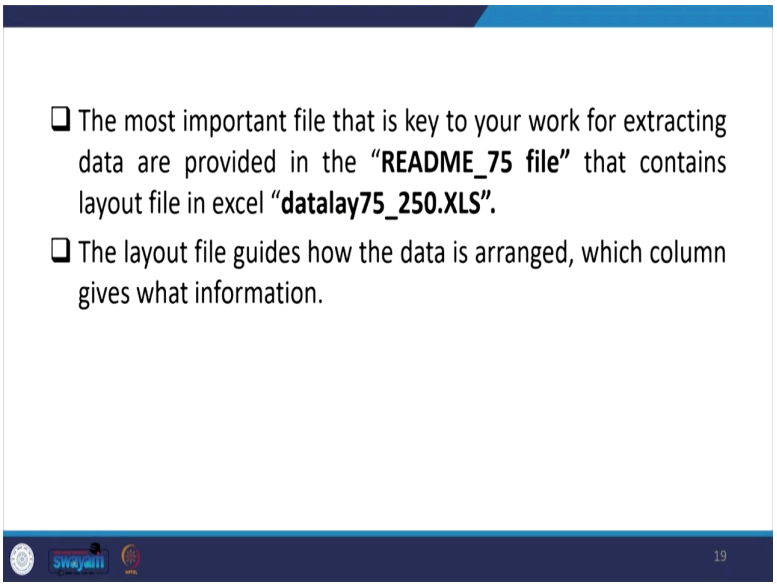
After saying so, I think we need to move it to the respective files. Once you have downloaded the raw data which from the previous page I have already shown you, once you keep your cursor on each of the txt file you can able to download. These data files are in ASCII format and downloaded from Readme file of 75th round. And the NSS 75th round data have separate data files for each level. So, for each 13 level we have separate data files.

(Refer Slide Time: 23:56)

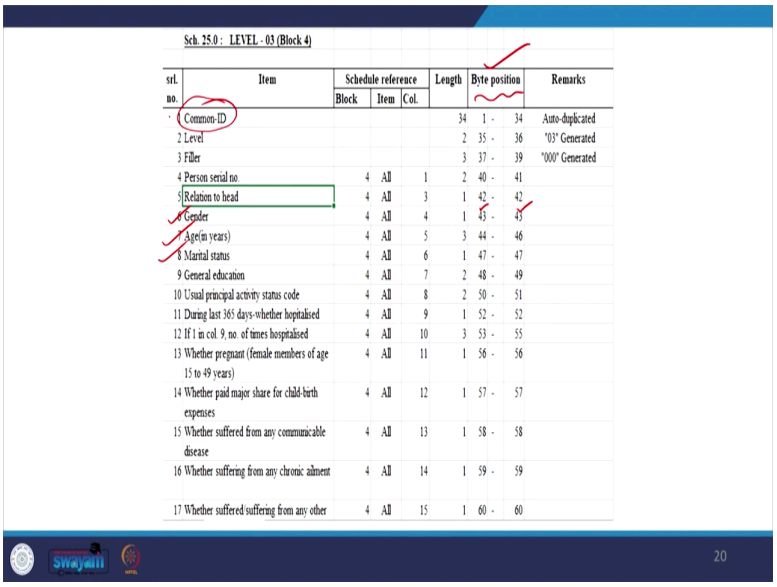


So, how it looks like? I will show everything once again with the original data. But at this moment I am just showing you the sample of it snapshot of that raw data. While each row in the data set represents a single observation with varying information about each unit without extraction it is impossible to do any meaningful analysis on these data sets. So, without reading between these things it is very difficult to understand in fact.

(Refer Slide Time: 23:36)



(Refer Slide Time: 23:38)

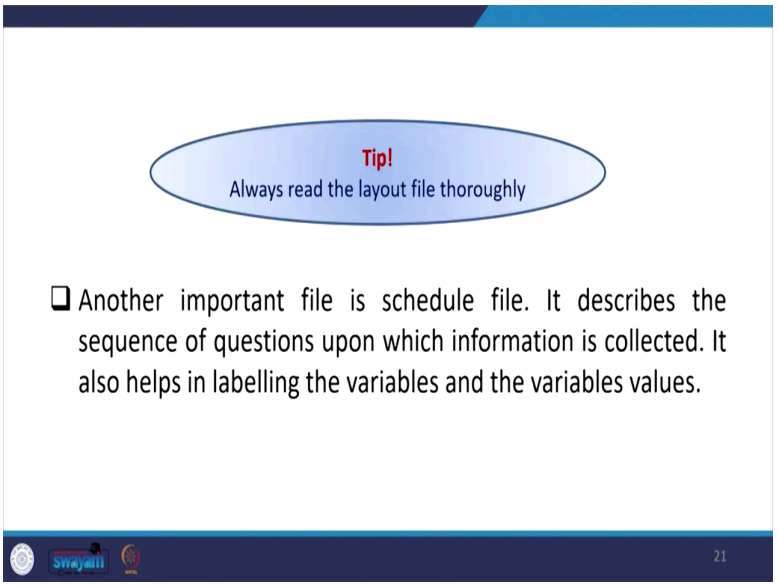


So, let me just guide you some other important aspects. The most important file that is key to your work or to your extraction is the Readme file and the layout file. The layout file guides how the data is arranged and in which column you get the variable and their information. This is how the layout file looks like. This is how it is visible.

So, like for example, you asked about individual file about the gender, their age, their marital status, their relation to head, their byte positions are important. You have to look at these bytes position like gender as we know that it would be either 1, 2 or 3. So, at maximum 1 byte space it will consume. So, byte space is 43 to 43; that means, on the data they have entered they have only given 1 digit space; that means, at maximum it can take value from 0 till 9.

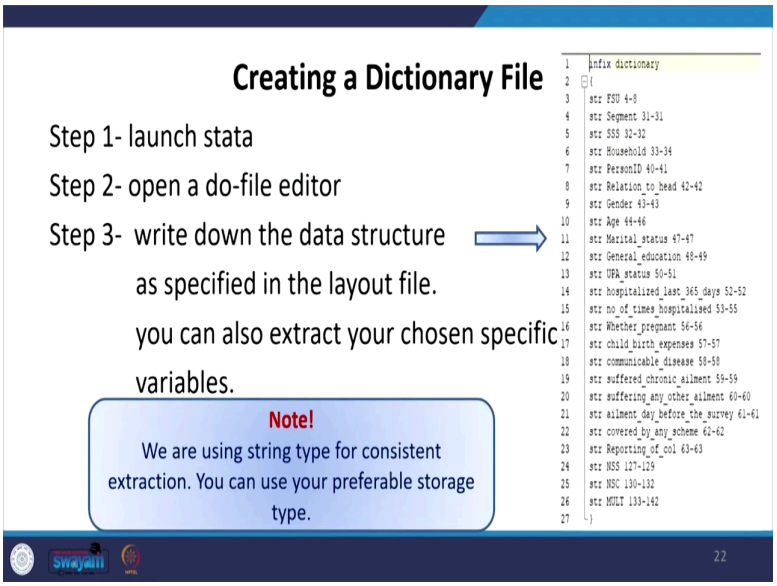
So, then similarly others other entries you can able to read. We are now just going to guide you about common ID that is quite important for merging and we will explain you in detail.

(Refer Slide Time: 24:57)



Maybe in the next last we will clarify very clearly about common ID. At this moment we are trying to extract it. One of the tips is given here that we need to always read the layout file very thoroughly. Another important file is called schedule file that is called the questionnaire file. It describes the sequence of question upon which information is collected. It also helps in labelling the variables and the variable values.

(Refer Slide Time: 25:29)

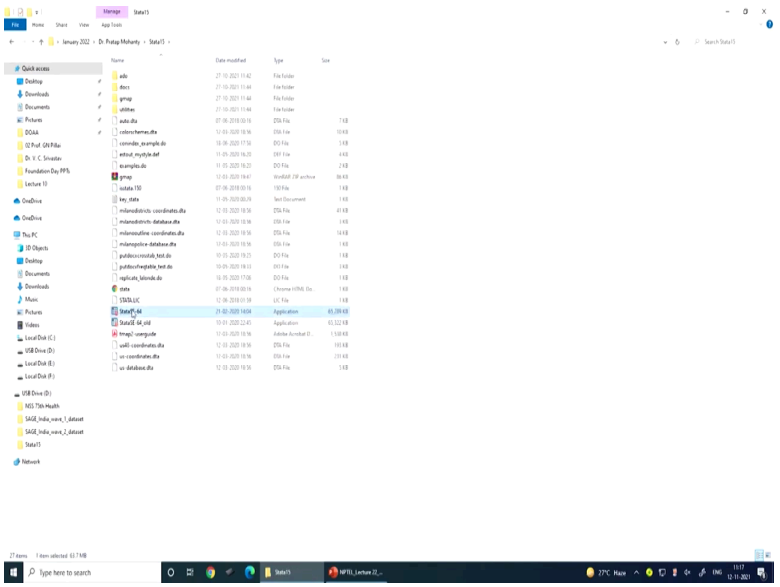


Then we will guide you once you have done those basic check or the basic files next aspect for you to check with the dictionary file. You have to create a dictionary file if you are

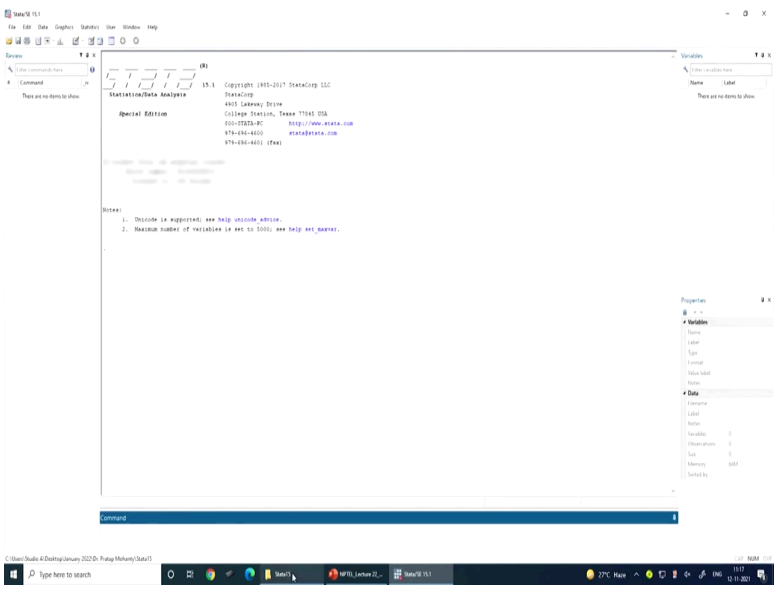
handling a bigger database, without dictionary file you can also able to extract data but what we are suggesting you that it will minimize your time and your result will be always be more correct if you follow an appropriate dictionary file. Then how to go for it for creating a dictionary file? There are different methods of extraction, I am going to discuss about it.

But, we can now open the data set and I will show you both together. I will guide you how to extract and which method we can apply. And I will then do simultaneous attempts to clarify all those details.

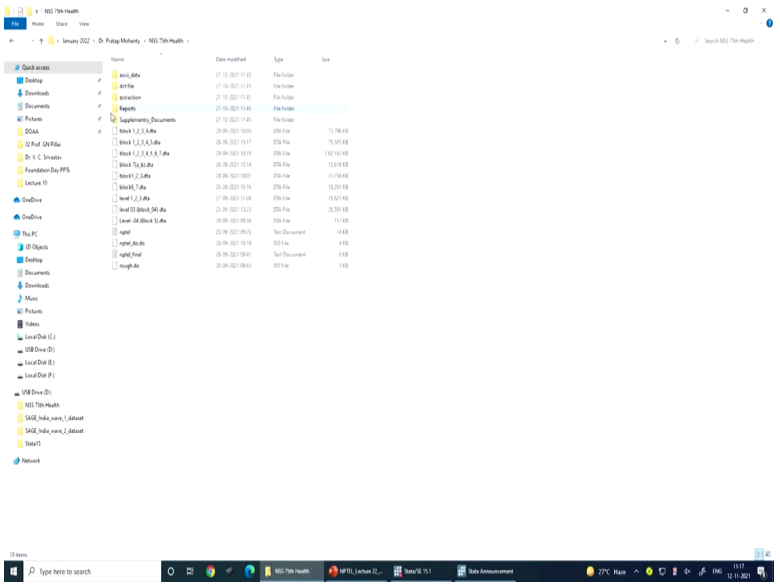
(Refer Slide Time: 26:33)



(Refer Slide Time: 26:39)



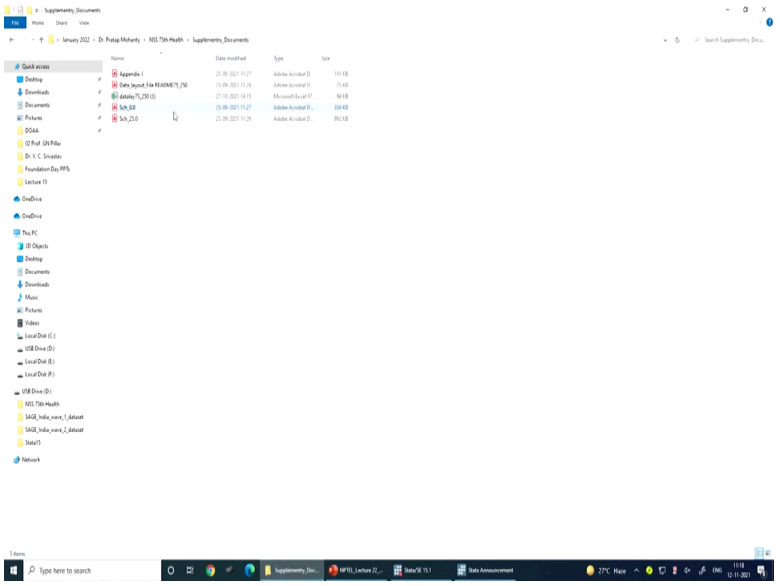
(Refer Slide Time: 26:46)



So, first of all, here is 75th round. We are also opening a Stata window here and we are also opening NSS 75th on health. And the PPT you can also open simultaneously will guide you all those thing side by side.

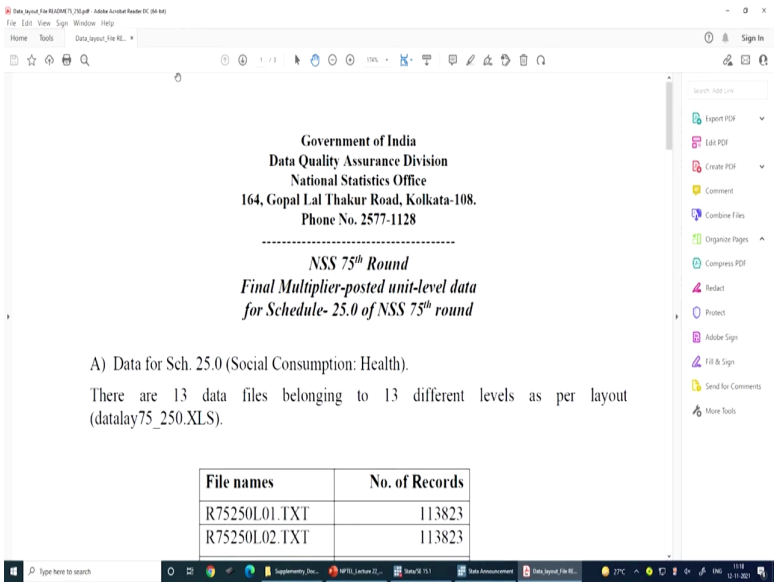
So, let me go quickly to the respective page which I have already said. So, first what I do? I will open and show you what is this data file, the important files which we have been discussing.

(Refer Slide Time: 27:09)



First of all, the supplementary document which I have said you have to open and read it very carefully.

(Refer Slide Time: 27:24)



Government of India
Data Quality Assurance Division
National Statistics Office
164, Gopal Lal Thakur Road, Kolkata-108.
Phone No. 2577-1128

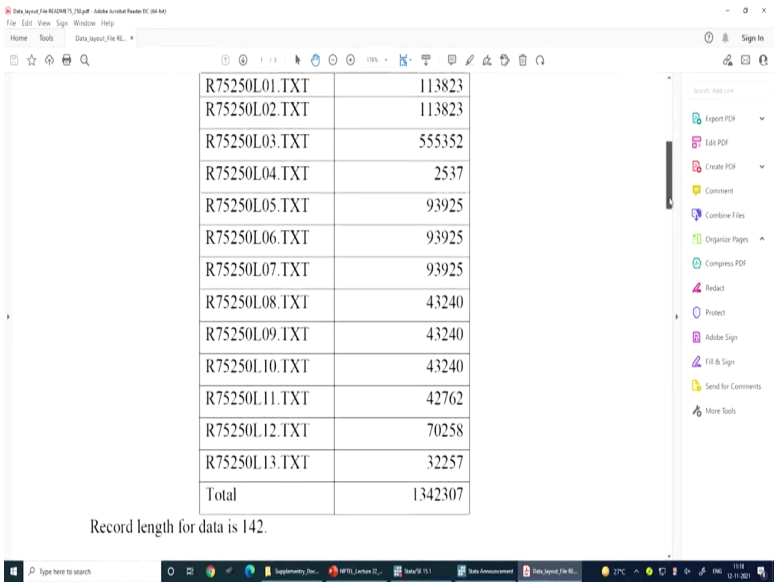
NSS 75th Round
Final Multiplier-posted unit-level data
for Schedule- 25.0 of NSS 75th round

A) Data for Sch. 25.0 (Social Consumption: Health).

There are 13 data files belonging to 13 different levels as per layout (datalay75_250.XLS).

File names	No. of Records
R75250L01.TXT	113823
R75250L02.TXT	113823

(Refer Slide Time: 27:29)

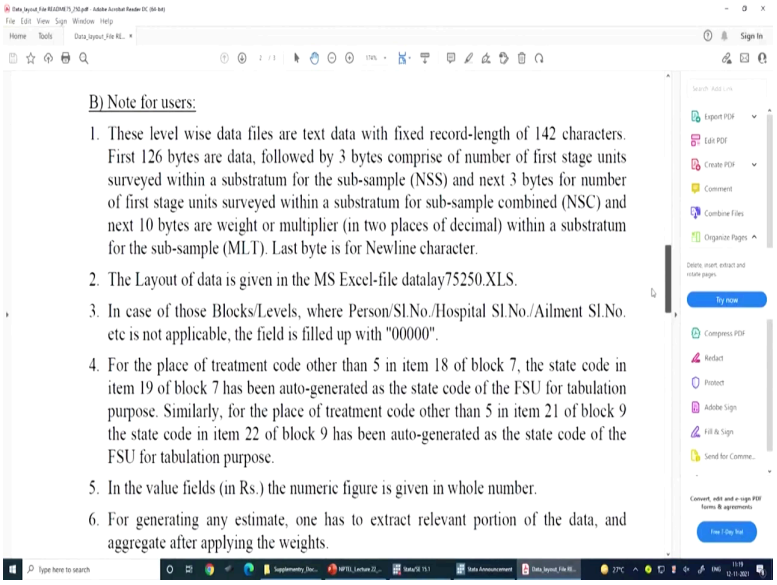


R75250L01.TXT	113823
R75250L02.TXT	113823
R75250L03.TXT	555352
R75250L04.TXT	2537
R75250L05.TXT	93925
R75250L06.TXT	93925
R75250L07.TXT	93925
R75250L08.TXT	43240
R75250L09.TXT	43240
R75250L10.TXT	43240
R75250L11.TXT	42762
R75250L12.TXT	70258
R75250L13.TXT	32257
Total	1342307

Record length for data is 142.

README file I have just opened on the screen. This looks like the one which we have shown it on the PPT. They are the different text, the layout file is available here, once I take my cursor it gives me the thumb to click on it and you can download it. So, this is the one you can download. And these are all file, if you keep your cursor over here, it asks for downloading.

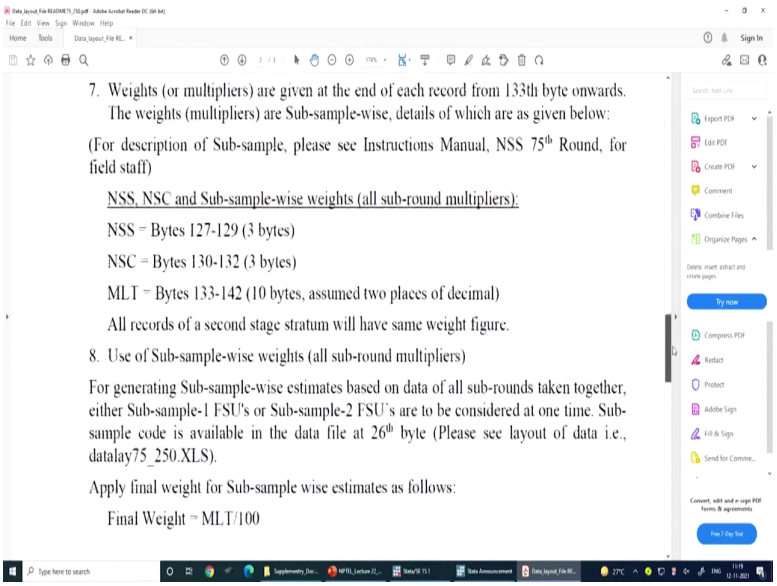
(Refer Slide Time: 27:58)



B) Note for users:

1. These level wise data files are text data with fixed record-length of 142 characters. First 126 bytes are data, followed by 3 bytes comprise of number of first stage units surveyed within a substratum for the sub-sample (NSS) and next 3 bytes for number of first stage units surveyed within a substratum for sub-sample combined (NSC) and next 10 bytes are weight or multiplier (in two places of decimal) within a substratum for the sub-sample (MLT). Last byte is for Newline character.
2. The Layout of data is given in the MS Excel-file datalay75250.XLS.
3. In case of those Blocks/Levels, where Person/Sl.No./Hospital Sl.No./Ailment Sl.No. etc is not applicable, the field is filled up with "00000".
4. For the place of treatment code other than 5 in item 18 of block 7, the state code in item 19 of block 7 has been auto-generated as the state code of the FSU for tabulation purpose. Similarly, for the place of treatment code other than 5 in item 21 of block 9 the state code in item 22 of block 9 has been auto-generated as the state code of the FSU for tabulation purpose.
5. In the value fields (in Rs.) the numeric figure is given in whole number.
6. For generating any estimate, one has to extract relevant portion of the data, and aggregate after applying the weights.

(Refer Slide Time: 28:08)



7. Weights (or multipliers) are given at the end of each record from 133th byte onwards. The weights (multipliers) are Sub-sample-wise, details of which are as given below:

(For description of Sub-sample, please see Instructions Manual, NSS 75th Round, for field staff)

NSS, NSC and Sub-sample-wise weights (all sub-round multipliers):

NSS = Bytes 127-129 (3 bytes)

NSC = Bytes 130-132 (3 bytes)

MLT = Bytes 133-142 (10 bytes, assumed two places of decimal)

All records of a second stage stratum will have same weight figure.

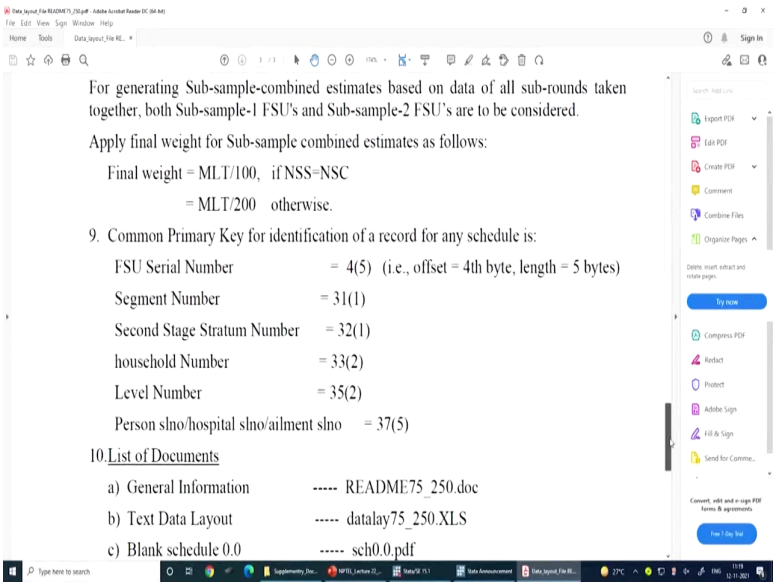
8. Use of Sub-sample-wise weights (all sub-round multipliers)

For generating Sub-sample-wise estimates based on data of all sub-rounds taken together, either Sub-sample-1 FSU's or Sub-sample-2 FSU's are to be considered at one time. Sub-sample code is available in the data file at 26th byte (Please see layout of data i.e., datalay75_250.XLS).

Apply final weight for Sub-sample wise estimates as follows:

Final Weight = MLT/100

(Refer Slide Time: 28:18)

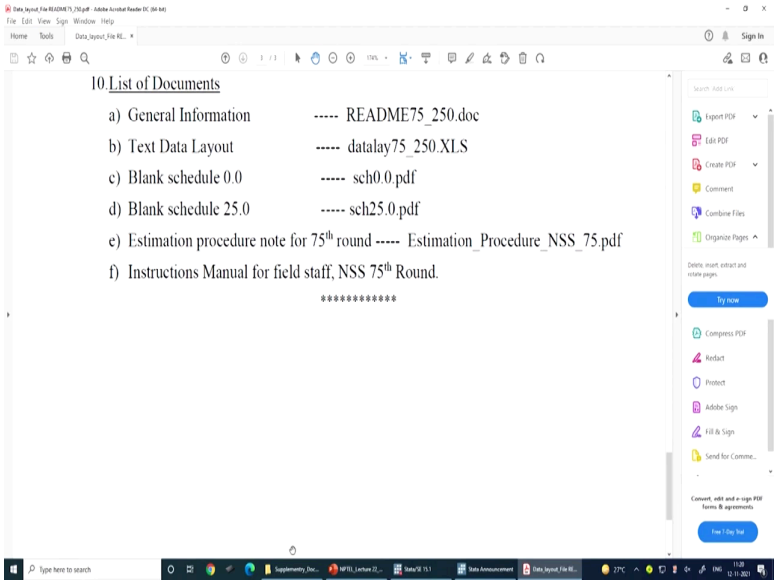


Then other important aspect you must read that is, they have guided about multiplier. A multiplier in this case is considered to be the weight for the analysis; how the weight is defined what do you mean by final weight. And they have guided you about the final weight here they multiplied divided by 100 if NSS and NSC are equal otherwise it is multiplied divided by 200.

And what is this multiplier file is already defined on the data I will show it on the layout file. The next important information here is to know about the common primary ID. I will also emphasize this several times. The common primary ID as per the suggestions in the data are consisting of these 6 important indicators

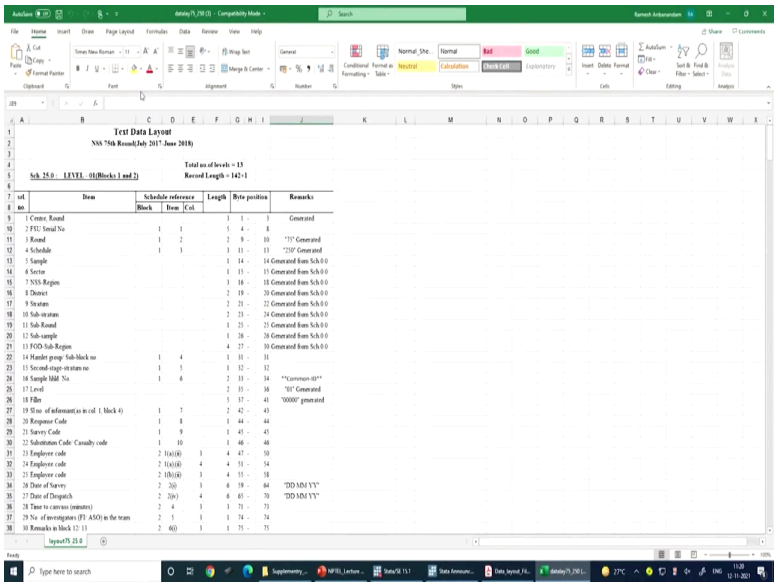
FSU, then segment number then second stage stratum number, household number, level number then person ID or element number serial number. So, if you are working of on the personal ID or on the individual file then you may be concerned for this particular position.

(Refer Slide Time: 29:17)



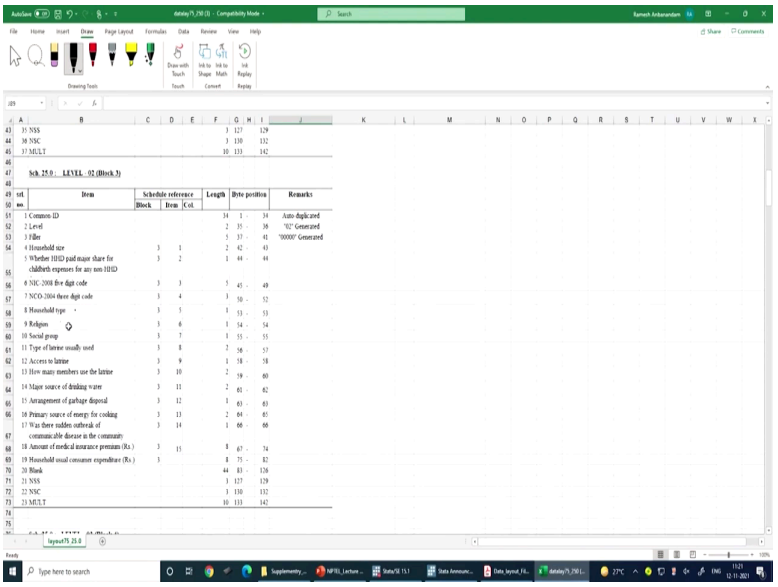
Then other important documents they have also suggested us to read. So, we are going to read all those things like your schedule, your layout file etc.

(Refer Slide Time: 29:48)



I am going to open the layout file then just to show you the exact file and for your understanding. So, this is in Excel format and this gives information about byte positions and the column wise position is given. I have just opened it for your reference.

(Refer Slide Time: 30:20)

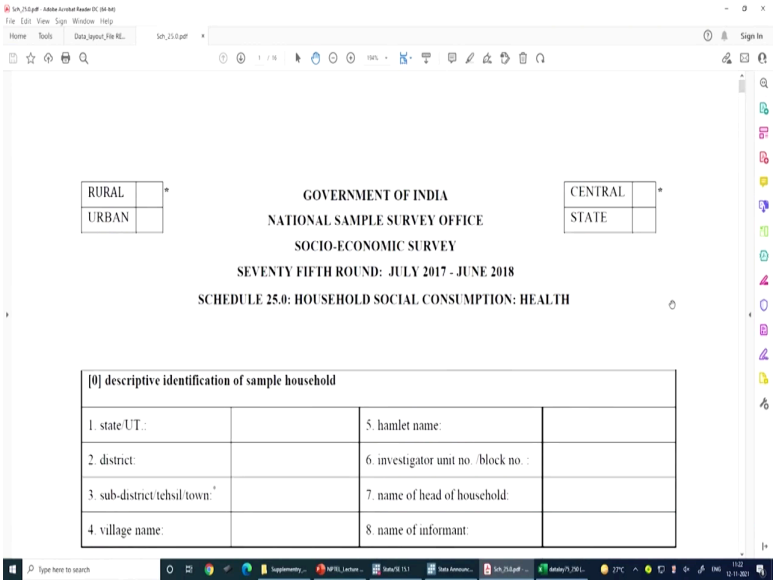


The first level is going to give us about the server related information its records. Like records on sampling related information, digit related information, round related information, district related information. So, this one is actually help us to understand the primary key.

From second onwards it is very important. Second level is for the household. Here, household type, then religion then social group then household size etc. and their byte positions are given. Byte positions like it is here, suppose you are saying household type, it has the byte position from 53 to 53, then religion it has the position 54 to 54.

So, these are important. I will also guide you about schedule. So, as I already told you during the understanding about data schedule 25 is meant for health care in all the rounds they have defined schedule 25 for health care.

(Refer Slide Time: 31:02)



RURAL		*
URBAN		

GOVERNMENT OF INDIA

NATIONAL SAMPLE SURVEY OFFICE

SOCIO-ECONOMIC SURVEY

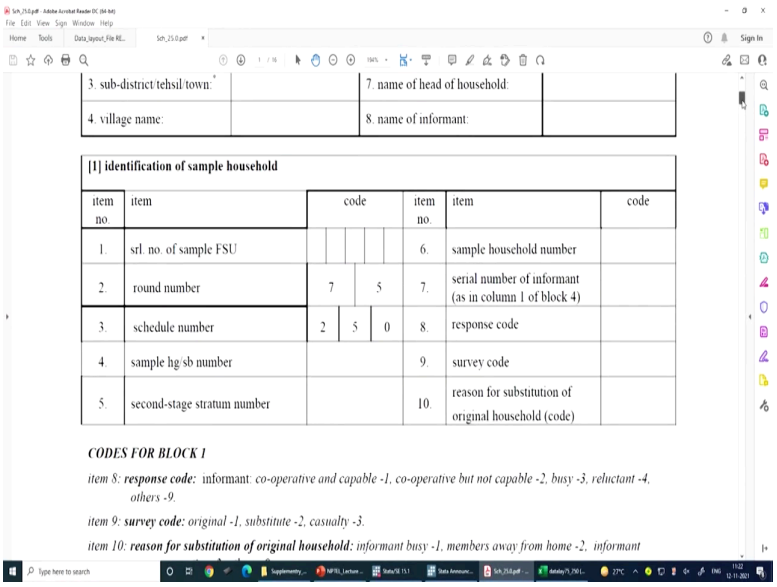
SEVENTY FIFTH ROUND: JULY 2017 - JUNE 2018

SCHEDULE 25.0: HOUSEHOLD SOCIAL CONSUMPTION: HEALTH

CENTRAL		*
STATE		

[0] descriptive identification of sample household			
1. state/UT.:		5. hamlet name:	
2. district:		6. investigator unit no. /block no. :	
3. sub-district/tehsil/town: *		7. name of head of household:	
4. village name:		8. name of informant:	

(Refer Slide Time: 31:11)



3. sub-district/tehsil/town: *		7. name of head of household:	
4. village name:		8. name of informant:	

[1] identification of sample household

item no.	item	code	item no.	item	code
1.	srl. no. of sample FSU		6.	sample household number	
2.	round number	7 5	7.	serial number of informant (as in column 1 of block 4)	
3.	schedule number	2 5 0	8.	response code	
4	sample hg/sb number		9.	survey code	
5.	second-stage stratum number		10.	reason for substitution of original household (code)	

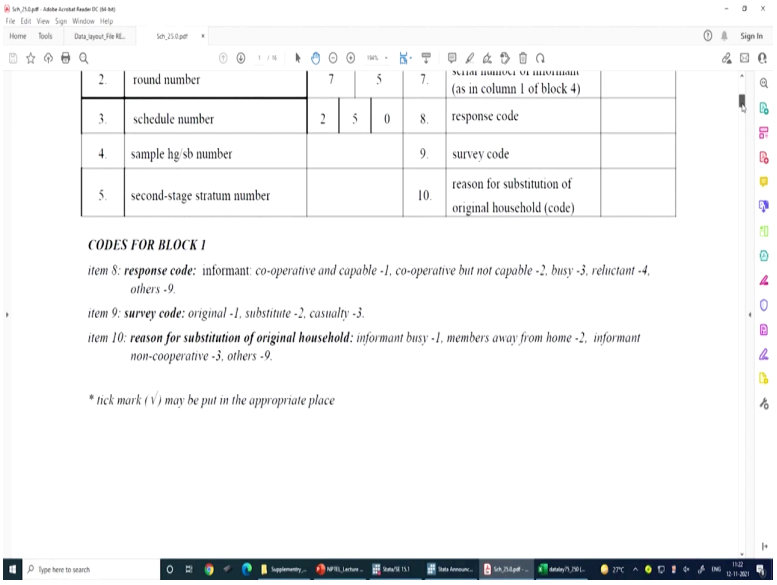
CODES FOR BLOCK 1

item 8: ***response code:*** informant: *co-operative and capable -1, co-operative but not capable -2, busy -3, reluctant -4, others -9.*

item 9: ***survey code:*** *original -1, substitute -2, casualty -3.*

item 10: ***reason for substitution of original household:*** *informant busy -1, members away from home -2, informant*

(Refer Slide Time: 31:13)



2.	round number	7	5	7.		(as in column 1 of block 4)	
3.	schedule number	2	5	0	8.	response code	
4.	sample hg/sb number				9.	survey code	
5.	second-stage stratum number				10.	reason for substitution of original household (code)	

CODES FOR BLOCK 1

item 8: ***response code:*** informant: *co-operative and capable -1, co-operative but not capable -2, busy -3, reluctant -4, others -9.*

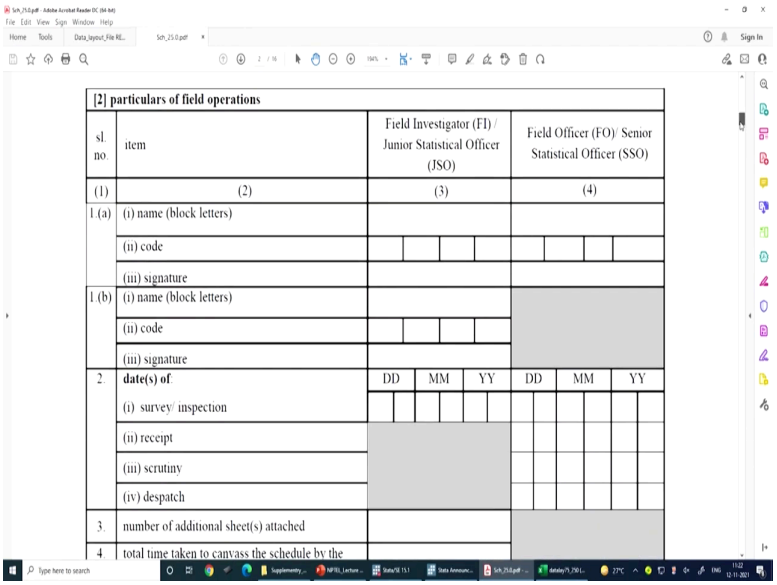
item 9: ***survey code:*** *original -1, substitute -2, casualty -3.*

item 10: ***reason for substitution of original household:*** *informant busy -1, members away from home -2, informant non-cooperative -3, others -9.*

** tick mark (√) may be put in the appropriate place*

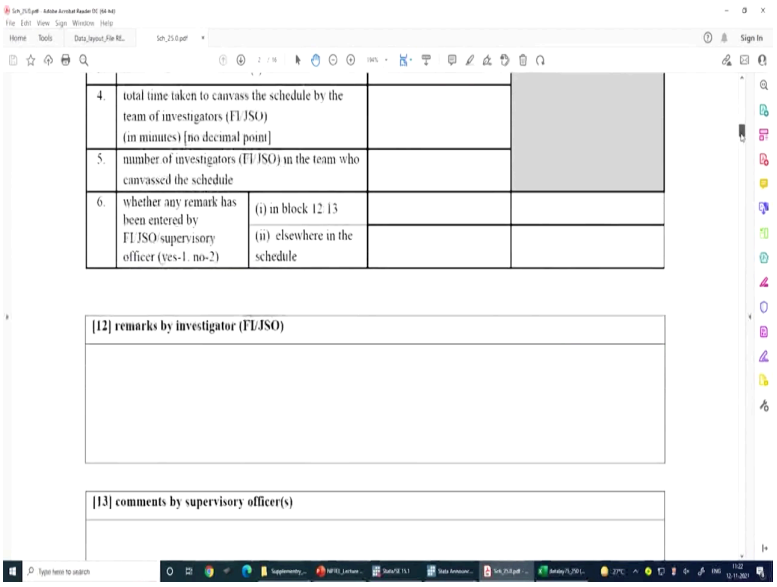
Then this is the questionnaire or this is the schedule where you exactly going to get it. On the very first identification related information available in our first layout or the and the layout on the first block of information.

(Refer Slide Time: 31:22)



[2] particulars of field operations								
sl. no.	item	Field Investigator (FI) / Junior Statistical Officer (JSO)			Field Officer (FO)/ Senior Statistical Officer (SSO)			
(1)	(2)	(3)			(4)			
1.(a)	(i) name (block letters)							
	(ii) code							
	(iii) signature							
1.(b)	(i) name (block letters)							
	(ii) code							
	(iii) signature							
2.	**date(s) of**:	DD	MM	YY	DD	MM	YY	
	(i) survey/ inspection							
	(ii) receipt							
	(iii) scrutiny							
	(iv) despatch							
3.	number of additional sheet(s) attached							
4.	total time taken to canvass the schedule by the							

(Refer Slide Time: 31:24)



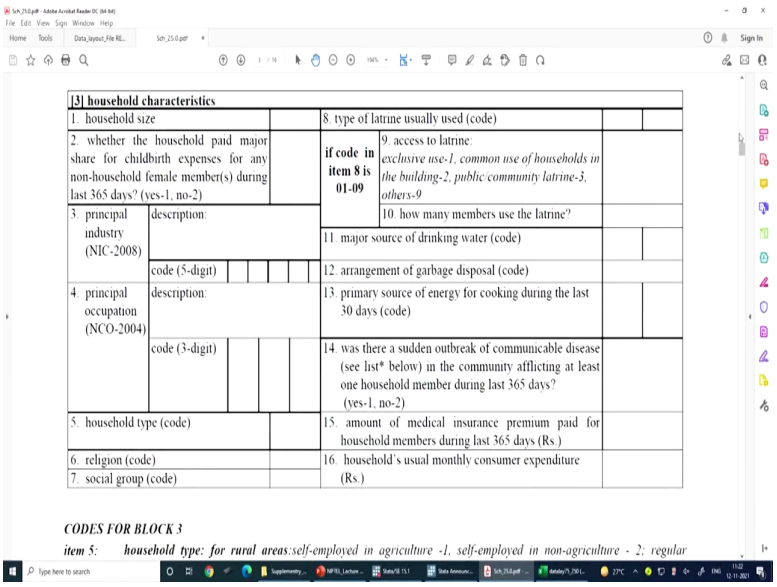
4.	total time taken to canvass the schedule by the team of investigators (FI/JSO) (in minutes) [no decimal point]			
5.	number of investigators (FI/JSO) in the team who canvassed the schedule			
6.	whether any remark has been entered by FI/JSO/supervisory officer (yes-1, no-2)	(i) in block 12 13		
		(ii) elsewhere in the schedule		

[12] remarks by investigator (FI/JSO)

[13] comments by supervisory officer(s)

From the second one it is all about the field operation from the third it is level third it is your household information.

(Refer Slide Time: 31:30)



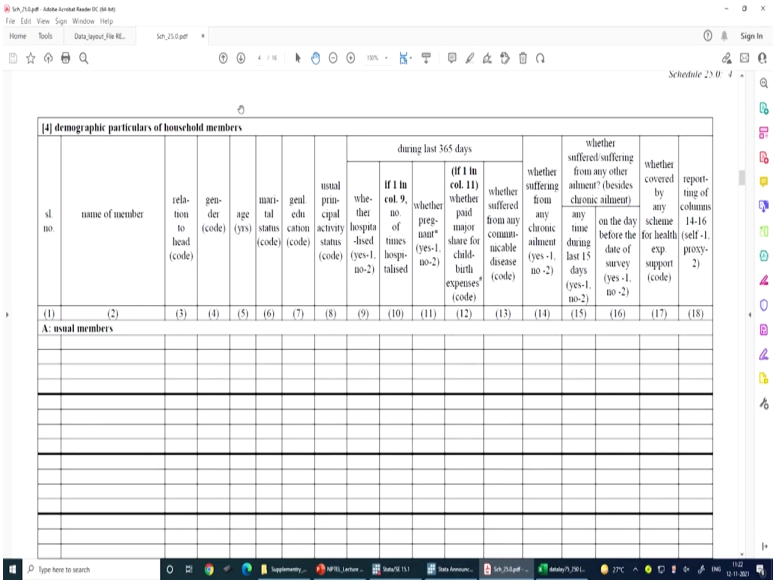
[3] household characteristics				
1. household size			8. type of latrine usually used (code)	
2. whether the household paid major share for childbirth expenses for any non-household female member(s) during last 365 days? (yes-1, no-2)		if code in item 8 is 01-09	9. access to latrine: *exclusive use-1, common use of households in the building-2, public/community latrine-3, others-9*	
3. principal industry (NIC-2008)	description:		10. how many members use the latrine?	
			11. major source of drinking water (code)	
	code (5-digit)		12. arrangement of garbage disposal (code)	
4. principal occupation (NCO-2004)	description:		13. primary source of energy for cooking during the last 30 days (code)	
	code (3-digit)		14. was there a sudden outbreak of communicable disease (see list* below) in the community afflicting at least one household member during last 365 days? (yes-1, no-2)	
5. household type (code)			15. amount of medical insurance premium paid for household members during last 365 days (Rs.)	
6. religion (code)			16. household's usual monthly consumer expenditure (Rs.)	
7. social group (code)				

CODES FOR BLOCK 3

item 5: ***household type: for rural areas****:self-employed in agriculture -1, self-employed in non-agriculture - 2; regular*

Household size, household type, religion etc. these are mentioned. Along with that this also gives information about their principal status, working in industries or their occupation, principal occupation information as well.

(Refer Slide Time: 31:53)



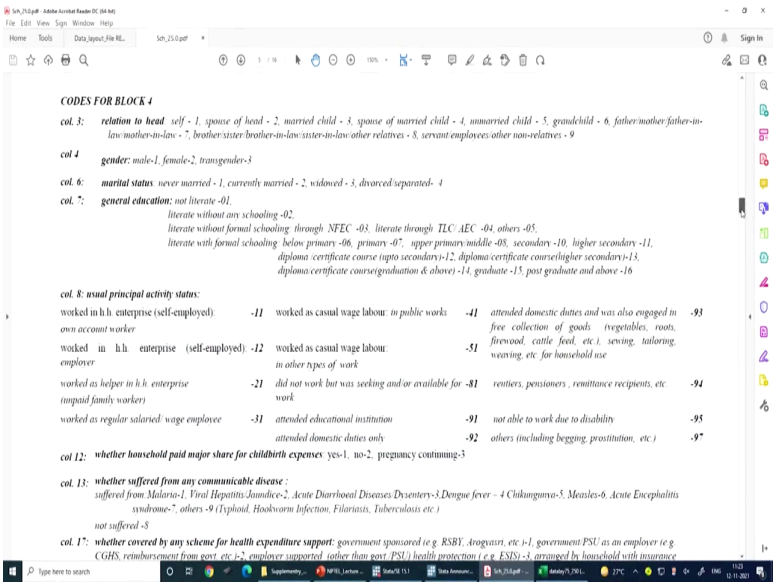
Schedule 25.0: 4

[4] demographic particulars of household members

sl. no.	name of member	rela-tion to head (code)	gen-der (code)	age (yrs)	mari-tal status (code)	geni. edu cation (code)	usual prin-cipal activity status (code)	during last 365 days: whe-ther hospita-lised (yes-1, no-2)	during last 365 days: if 1 in col. 9, no. of times hospi-talised	during last 365 days: whether preg-nant* (yes-1, no-2)	during last 365 days: (if 1 in col. 11) whether paid major share for child-birth expenses# (code)	during last 365 days: whether suffered from any comun-nicable disease (code)	whether suffering from any chronic ailment (yes -1, no -2)	whether suffered/suffering from any other ailment? (besides chronic ailment): any time during last 15 days (yes-1, no-2)	whether suffered/suffering from any other ailment? (besides chronic ailment): on the day before the date of survey (yes -1, no -2)	whether covered by any scheme for health exp. support (code)	report-ting of columns 14-16 (self -1, proxy-2)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
A: usual members																	

Next to this is your individual file like demographic particulars of household members. So, household member and their marital status, their age etc. is given.

(Refer Slide Time: 32:01)



CODES FOR BLOCK 4

col. 3: ***relation to head*** *self - 1, spouse of head - 2, married child - 3, spouse of married child - 4, unmarried child - 5, grandchild - 6, father/mother/father-in-law/mother-in-law - 7, brother/sister/brother-in-law/sister-in-law/other relatives - 8, servant/employees/other non-relatives - 9*

col 4 ***gender:*** *male-1, female-2, transgender-3*

col. 6: ***marital status*** *never married - 1, currently married - 2, widowed - 3, divorced/separated- 4*

col. 7: ***general education:*** *not literate -01,*
literate without any schooling -02,
literate without formal schooling through NFEC -03, literate through TLC/ AEC -04, others -05,
literate with formal schooling: below primary -06, primary -07, upper primary/middle -08, secondary -10, higher secondary -11, diploma /certificate course (upto secondary)-12, diploma/certificate course(higher secondary)-13, diploma/certificate course(graduation & above) -14, graduate -15, post graduate and above -16

col. 8: usual principal activity status:

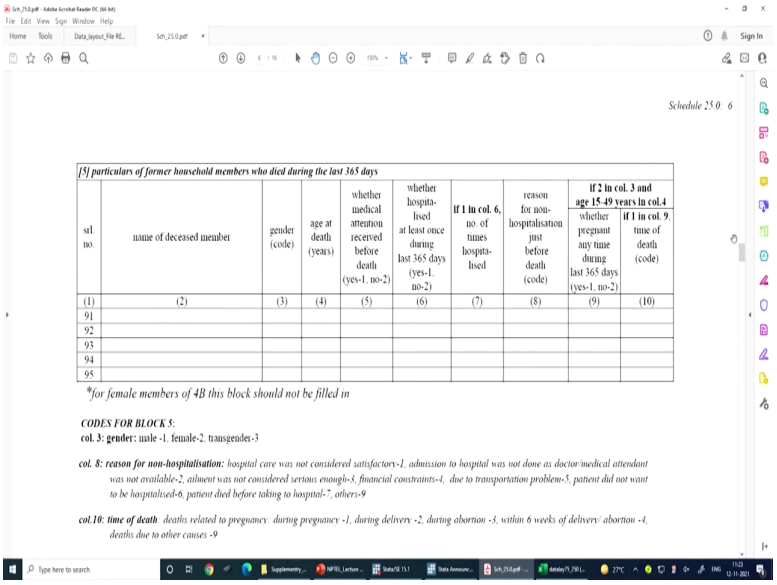
worked in h.h. enterprise (self-employed): *own account worker*	-11	worked as casual wage labour: *in public works*	-41	*attended domestic duties and was also engaged in free collection of goods (vegetables, roots, firewood, cattle feed, etc.), sewing, tailoring, weaving, etc. for household use*	-93
worked in h.h. enterprise (self-employed): *employer*	-12	worked as casual wage labour: *in other types of work*	-51		
worked as helper in h.h. enterprise (unpaid family worker)	-21	*did not work but was seeking and/or available for work*	-81	*rentiers, pensioners, remittance recipients, etc*	-94
worked as regular salaried/ wage employee	-31	*attended educational institution*	-91	*not able to work due to disability*	-95
		attended domestic duties only	-92	*others (including begging, prostitution, etc.)*	-97

col 12: ***whether household paid major share for childbirth expenses****:* yes-1, no-2, pregnancy continuing-3

col. 13: ***whether suffered from any communicable disease*** *:*
suffered from Malaria-1, Viral Hepatitis/Jaundice-2, Acute Diarrhoeal Diseases/Dysentery-3,Dengue fever – 4 Chikungunya-5, Measles-6, Acute Encephalitis syndrome-7, others -9 (Typhoid, Hookworm Infection, Filariasis, Tuberculosis etc.)
not suffered -8

col. 17: ***whether covered by any scheme for health expenditure support****: government sponsored (e.g. RSBY, Arogyasri, etc.)-1, government/PSU as an employer (e.g. CGHS, reimbursement from govt. etc.)-2, employer supported (other than govt./PSU) health protection (e.g. ESIS) -3, arranged by household with insurance*

(Refer Slide Time: 32:15)



Schedule 25.0: 6

[5] particulars of former household members who died during the last 365 days

srl no.	name of deceased member	gender (code)	age at death (years)	whether medical attention received before death (yes-1, no-2)	whether hospitalised at least once during last 365 days (yes-1, no-2)	If 1 in col. 6, no. of times hospitalised	reason for non-hospitalisation just before death (code)	If 2 in col. 3 and age 15-49 years in col.4	
								whether pregnant any time during last 365 days (yes-1, no-2)	If 1 in col. 9, time of death (code)
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
91									
92									
93									
94									
95									

*for female members of 4B this block should not be filled in

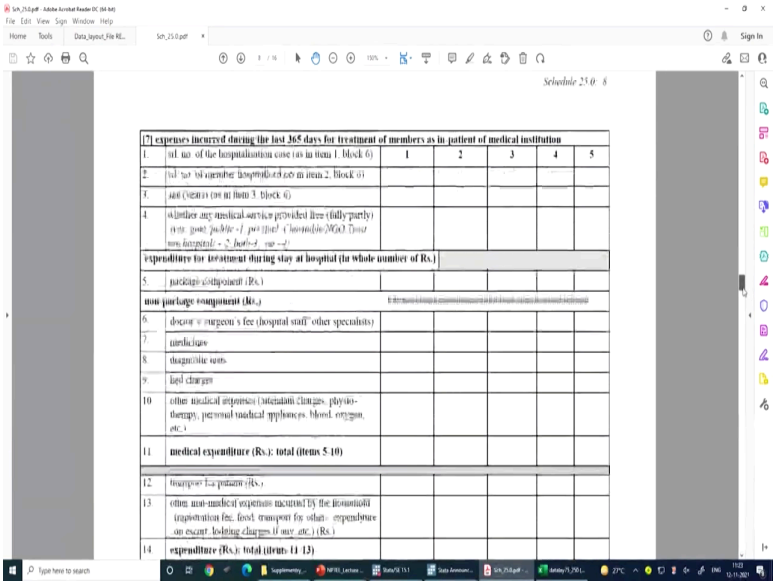
CODES FOR BLOCK 5:
col. 3: gender: male -1, female-2, transgender-3

col. 8: reason for non-hospitalisation: hospital care was not considered satisfactory-1, admission to hospital was not done as doctor/medical attendant was not available-2, ailment was not considered serious enough-3, financial constraints-4, due to transportation problem-5, patient did not want to be hospitalised-6, patient died before taking to hospital-7, others-9

col.10: time of death: deaths related to pregnancy: during pregnancy -1, during delivery -2, during abortion -3, within 6 weeks of delivery/ abortion -4, deaths due to other causes -9

And their different leveling, these codes are given, how these codes are entered in the text file is understood very clearly. Then your important aspect is to read, if you check with the 5th block here as per the schedule or in 4th block in the text file you have the information about the person who died during the last 365 days.

(Refer Slide Time: 32:36)

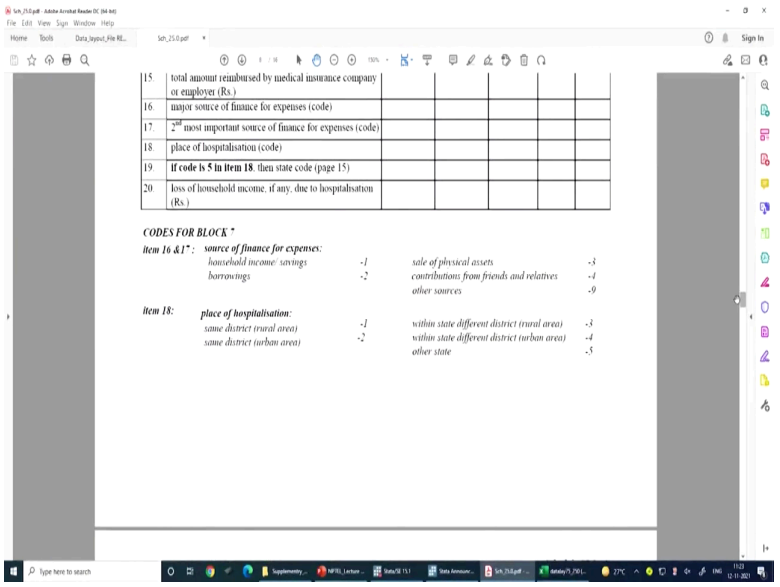


Schedule 25.0: 8

[7] expenses incurred during the last 365 days for treatment of members as in patient of medical institution

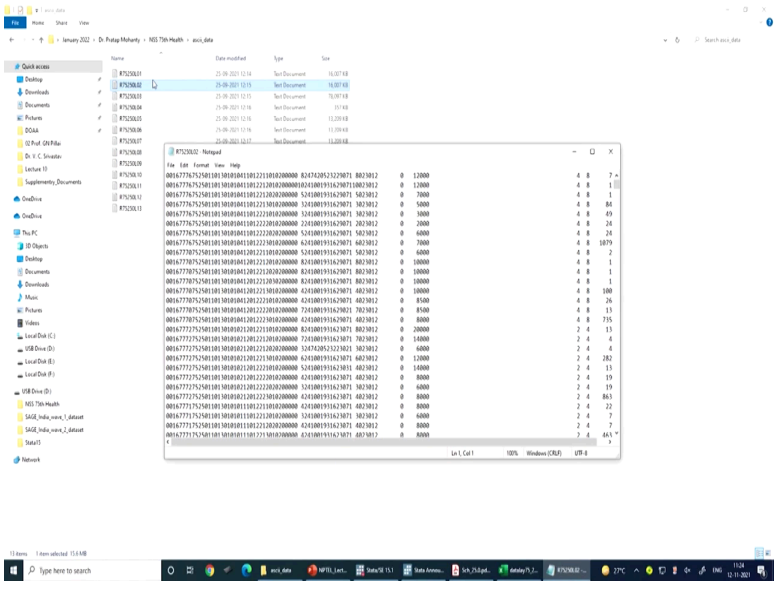
		1	2	3	4	5
1.	srl. no. of the hospitalisation case (as in item 1, block 6)					
2.	srl. no. of member hospitalised (as in item 2, block 6)					
3.	age (years) (as in item 3, block 6)					
4	whether any medical service provided free (fully/partly) (free govt. facility-1, partial-2, unavailable/NGO, free non-hospital-3, both-4, one-5)					
expenditure for treatment during stay at hospital (in whole number of Rs.)						
5.	package component (Rs.)					
non-package component (Rs.)						
6.	doctor's/surgeon's fee (hospital staff/ other specialists)					
7.	medicines					
8.	diagnostic tests					
9.	bed charges					
10	other medical expenses (anesthetic charges, physio-therapy, personal medical appliances, blood, oxygen, etc.)					
11	**medical expenditure (Rs.): total (items 5-10)**					
12.	transport [i.e. patient] (Rs.)					
13	other non-medical expenses incurred by the household (registration fee, food, transport for other expenditure on escort, lodging charges if any, etc.) (Rs.)					
14	**expenditure (Rs.): total (items 11-13)**					

(Refer Slide Time: 32:38)

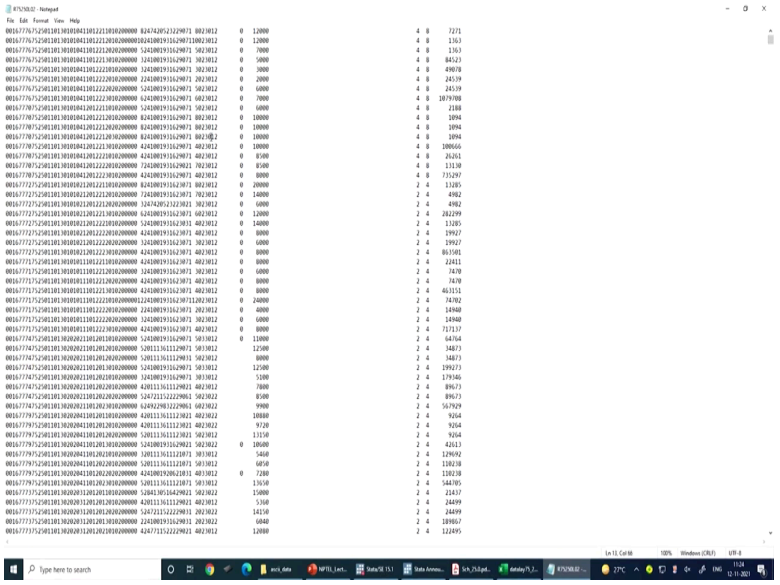


Similarly, next to this all information are related to individual file. I think we have done all those basic checks then we need to understand about ASCII data. ASCII data which I have already clarified, these are entered in the text file in the notepad file.

(Refer Slide Time: 33:01)



(Refer Slide Time: 33:18)



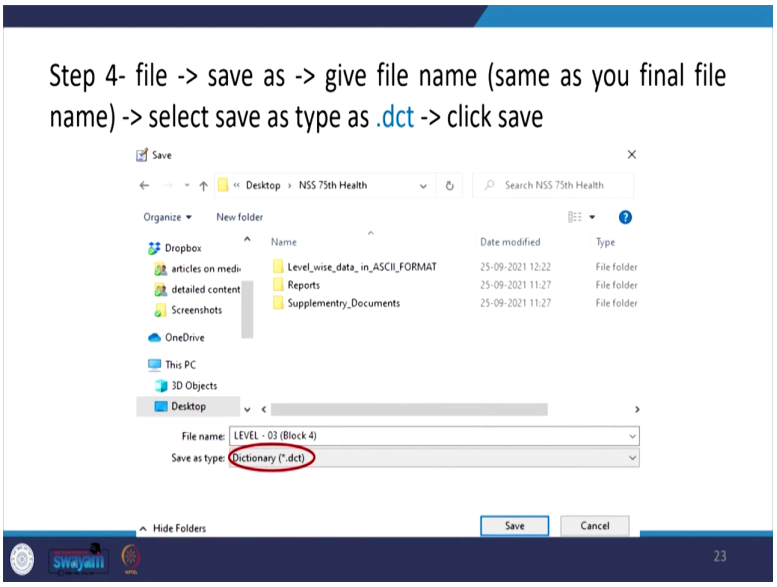
So, ASCII data like if I just open one file over here let it be these two. The sample which we have shown it is also visible here. How can you read it? Suppose we say that till the end, suppose I just keep my cursor over here, it will read your position. Your line number and it is column number can be shown at the bottom of this particular page.

Similarly, all its positions can be well read, like this is your first position, this is your second position, this is your third position etc. So, there are some spaces in given. Spaces you need not worry about it. In the data itself they have specified about how many positions the spaces have taken. So, we need not worry.

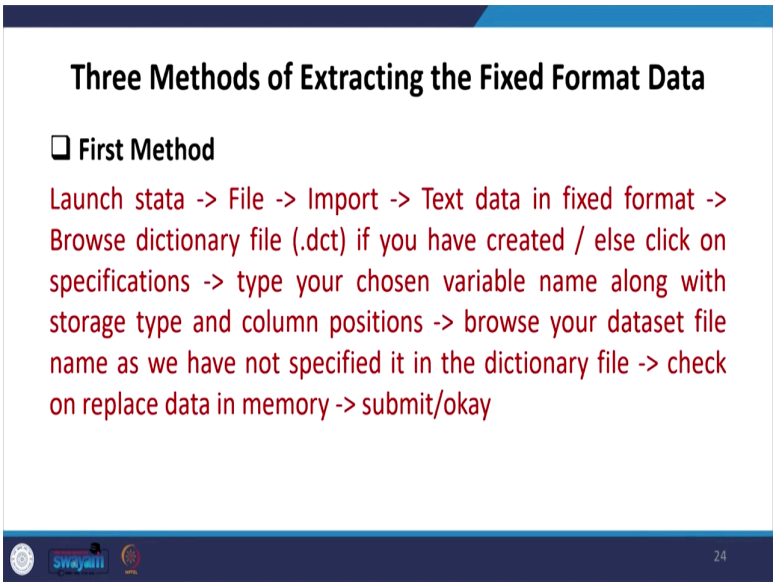
We need to read between the position byte position specified in the layout file. So, then the next aspect after guiding all those things I will go to your PPT once again. I have shown all those details. We need to create a dictionary file. So, now, we are handling the fixed format data, not free format, not delimited format.

We are dealing with the fixed form data where the fixed position is actually defined. So, its byte position is well defined in the layout file. Now, how we can create a dictionary file? There are different approaches to do it. But is it really very essential? You may not even require it, there are three methods where we can extract the data. One of the approaches is through the dct file.

(Refer Slide Time: 34:52)



(Refer Slide Time: 34:58)

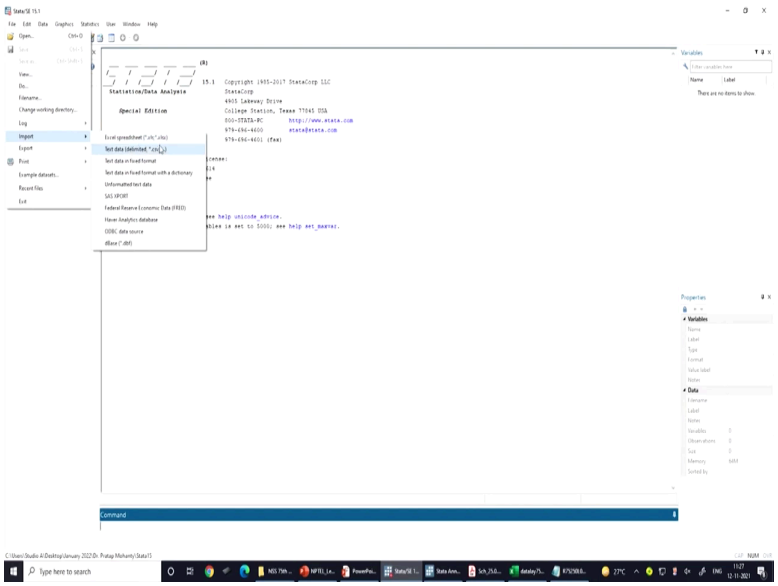


Let us start with a very basic of extraction. From the click-based approach we can able to extract the data as well. This is basically called the first method. I have also shown you some previous pages those who are already discussed. I am going to come back to the dictionary file once again.

Let me just stick to these first. Launching a Stata then we will go to File then Import then we will specify the dct file if it is defined. Otherwise, you can also specify the fixed format

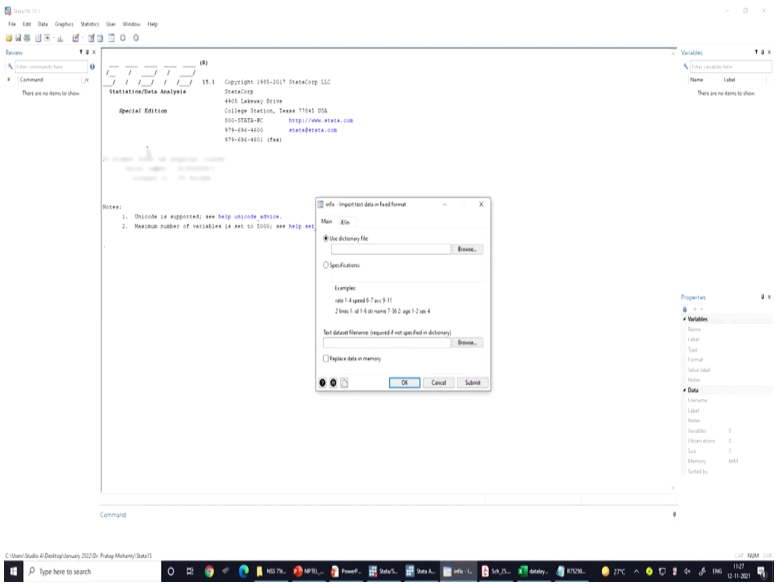
information then we can download. Like I can show it over here in your Stata window. Here we will go to File.

(Refer Slide Time: 35:56)



This is the first method then we will Import, but import from where? Import from the data in fixed format. I have said that our data is loaded with fixed format information. If it is on csv format or delimited format, then you could have been clicked here. If it is in an Excel format, but some specification needs to be very careful about it. Since it is in fixed format you have to simply click here.

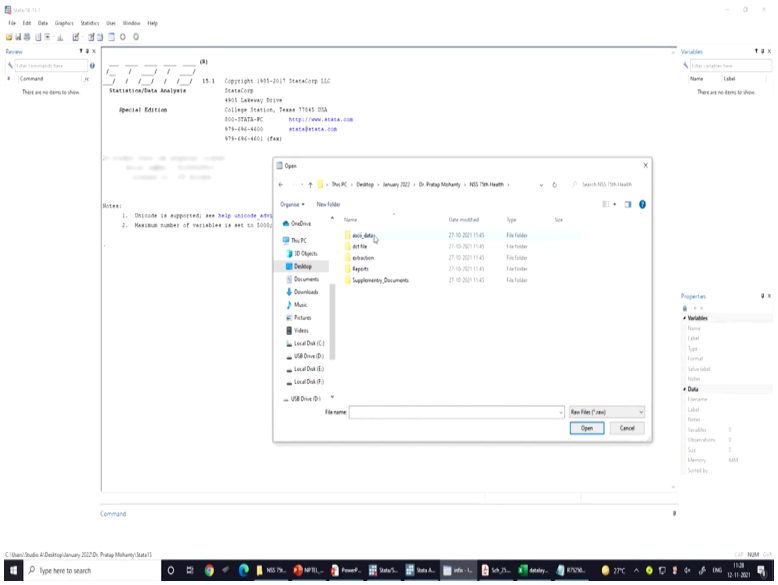
(Refer Slide Time: 36:21)



So, with this click it asks for a dictionary file. So, if your dictionary file has already been installed or you have already defined then you can open it. I will guide you how to open it and then you can operate it. Otherwise, you have to give the specifications. What do you mean by specification? You have to give the command which we are going to operate in our next page.

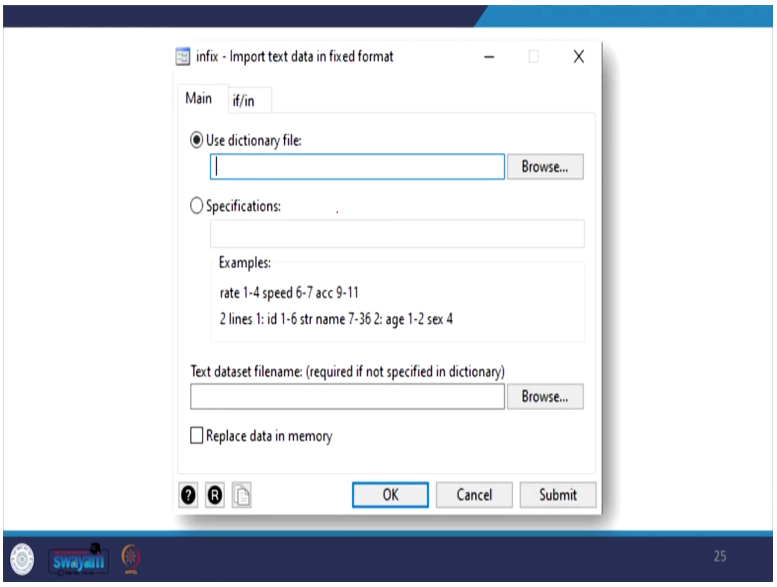
So, once we submit and end the source of the data. Here browsing means you are going to specify where the data we wanted to import, and your data has to be a micro data file.

(Refer Slide Time: 36:58)



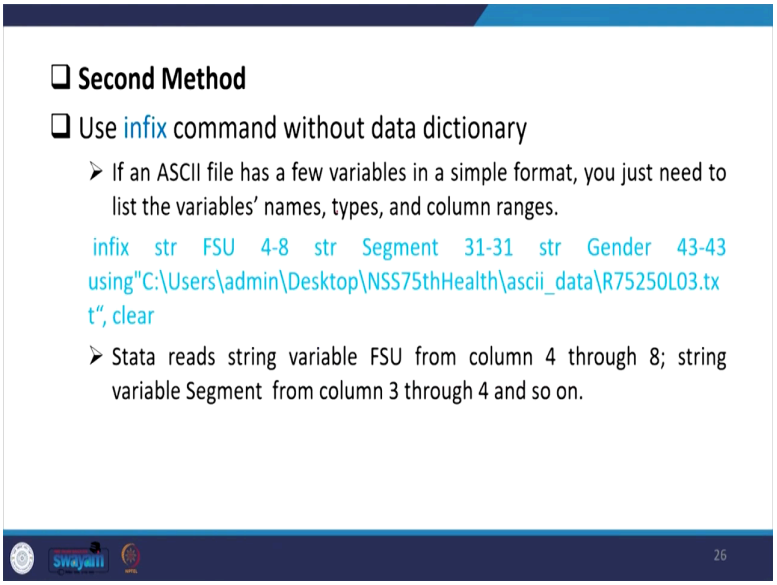
So, micro data file as I already told you it is ASCII data. So, here all files then all text files is visible and you can open it the respective page. So, let me just go back to our PPT once again and I will guide you rest of the details. This is one method.

(Refer Slide Time: 37:20)



This is how it looks like.

(Refer Slide Time: 37:21)

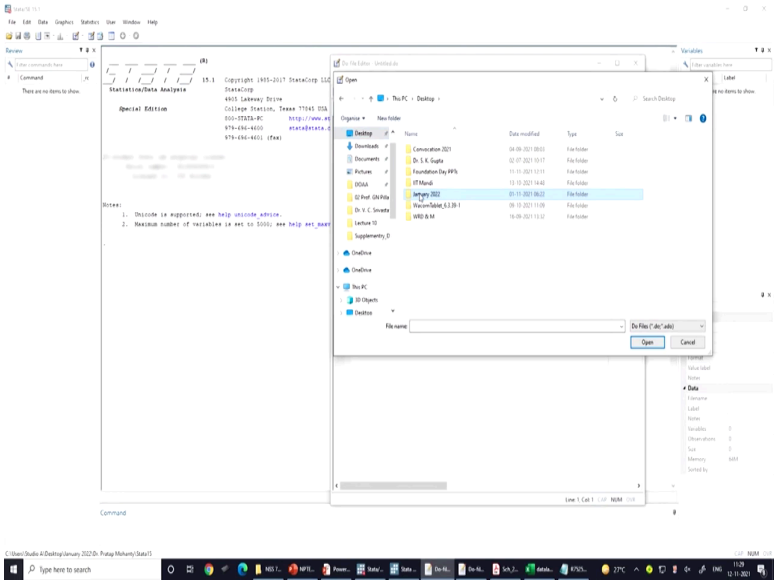


Then second method where we can go with the manual command. So, manual command on the command window or command space. So, we will have to give the infix command. infix command and its position. So, I am just going to show it to you here.

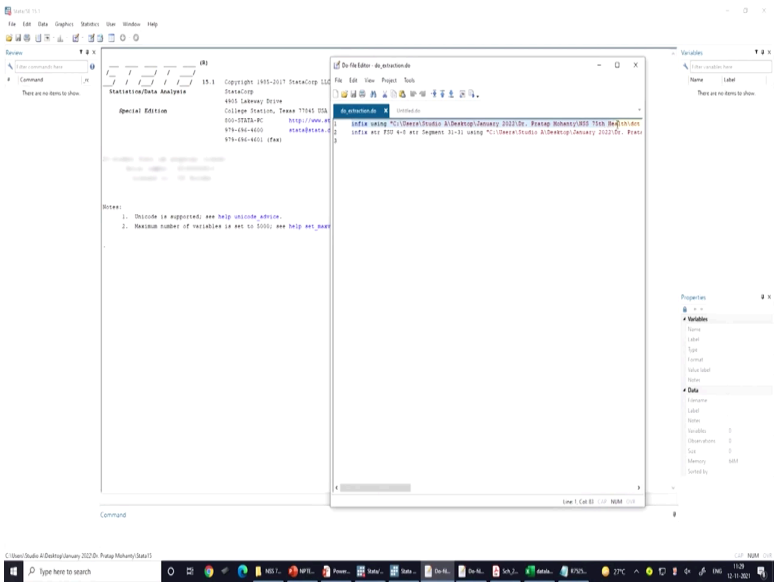
So, what we will do? We will first do couple of check like we can type the path name the way we have defined here. For the time being we are just going to open the command we have

saved it for your quick reference. You can easily type infix then using command, but we are going to show it how this looks like and how you can operate.

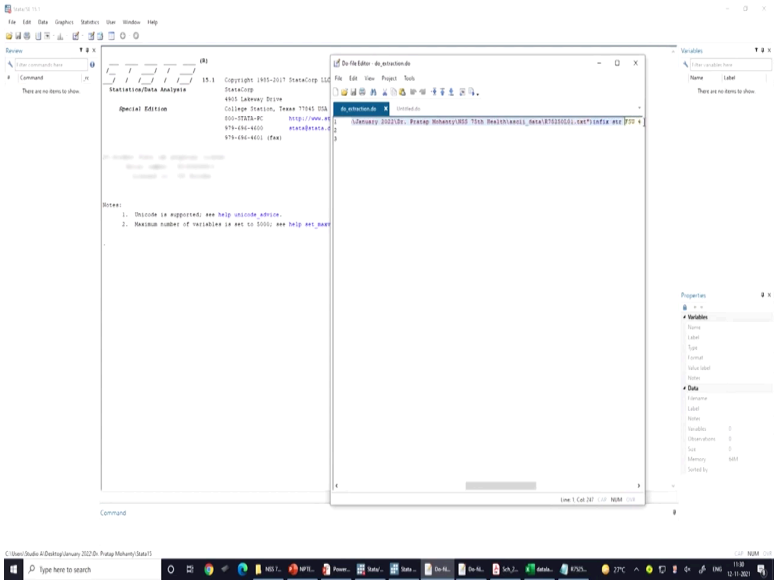
(Refer Slide Time: 38:21)



(Refer Slide Time: 38:33)



(Refer Slide Time: 38:40)

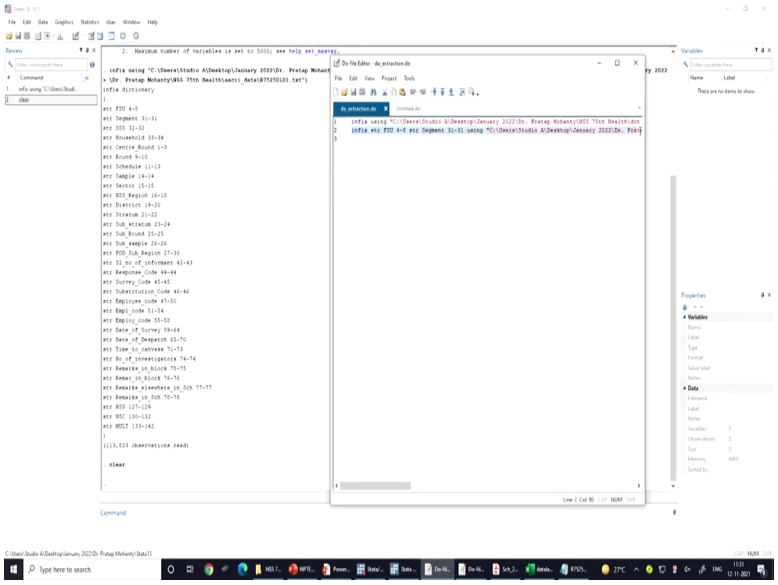


So, basically, we are guiding you on the path names we have already saved it for our reference. So, I am just going to show it over here. So, here it is and I will guide you what exactly you are supposed to write down. So, over here it you have to write down infix. Then using then on the using first party you are creating a dictionary file.

So, if you have already created a dictionary file then you could have returned the with the path, I will tell you what all about the dictionary file is. Then after comma it is your using file and them you have to browse raw data. So, the raw data should be ended with a dot txt.

And another sensitive aspect is that it should start with a inverted comma bracket and that has to be closed at the end. And another one is the backward slash has also to be given very correctly and the file path name has to be specified. Once we do it you can able to import the data.

(Refer Slide Time: 39:46)



We can just click on this and data will be extracted. Now, this is what is one of the methods where we have used the dct file. But if you do not use the dct file, you can simply type the command with space. Command with a byte position like infix you type infix another method which is manual method like infix we have to clear these first.

So, then clear the stored data then we type one by one; one by one you can just see these byte position in the layout file. Then you right in the command box like what it is written is infix. Then before using do remember that before using you have to mention the type of the data we are deliberately writing *str* stands for string.

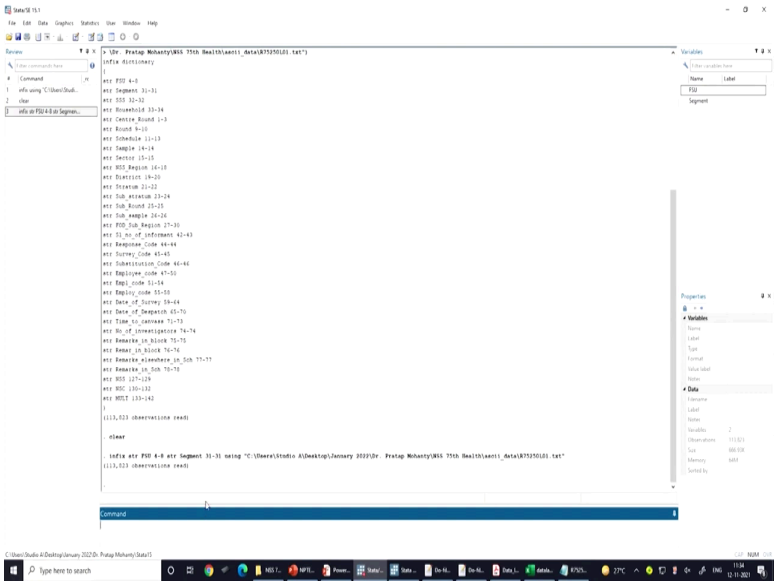
We are trying to make the data extracted in string format because that helps in not distributing the data that is compressing the data in string format and that further we can get the numeric form for further analysis. So, str for string we are making all the variables in string. At this moment we just wanted to show for two variables FSU and Segment and its position as I already told you FSU is from 4 to 8 that you can just check.

FSU is here 4 to 8. Position number 4 to 8 even it was guided on the serial number 2 or on the row number 10 of Excel sheet 4 to 8 is your FSU, then second one is your Segment is 31 to 31. So, this is what is given, 31 till 31 is your segment and that is also guided in the schedule file as well in the README file as well.

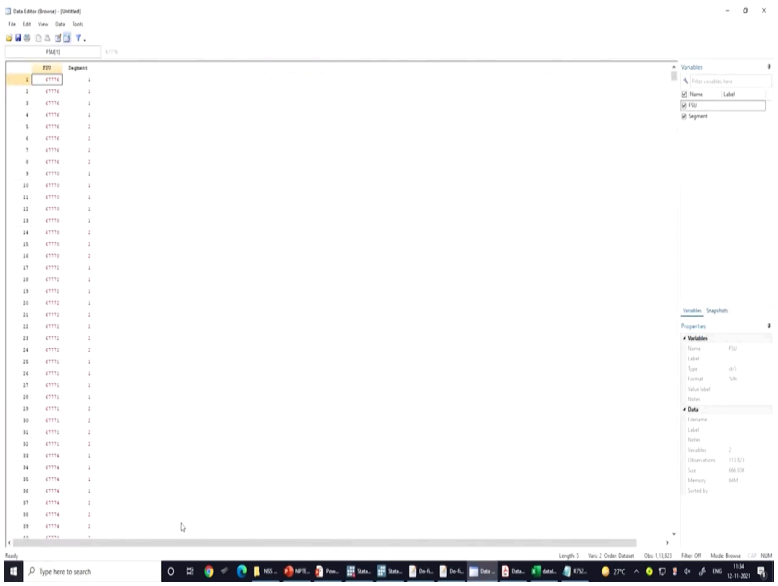
So, in the README file you have all such information. So, 31 to 31 is your segment number. Now after specifying this you can able to extract your data and make sure that you have used in after that using. If you have more variables then you can use all those variables over here.

So, this method is less prepared it this does not require a dictionary file. So, you have to do it manually. At the end you need to give the path name of this text file that is your raw data or the micro data dot txt has to be mentioned and inverted comma has to be closed. So, now, we can enter it and the data will be extracted.

(Refer Slide Time: 42:50)



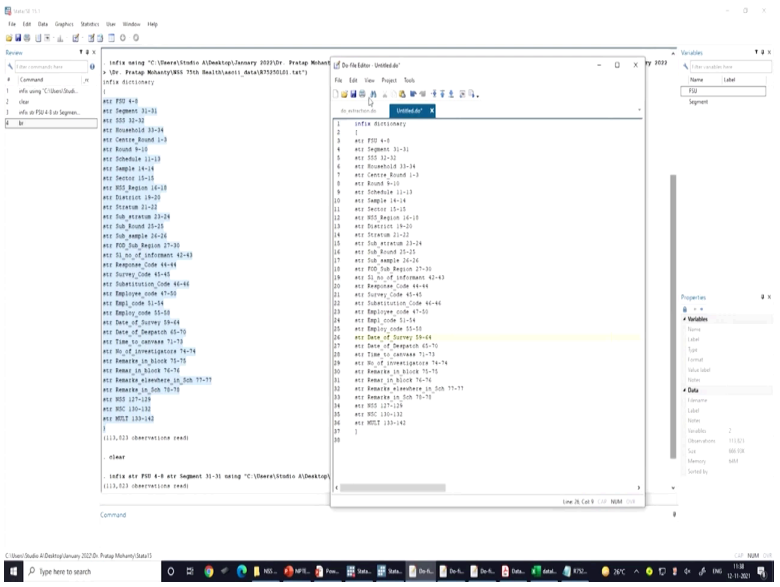
(Refer Slide Time: 42:56)



So, you can just see whether data is extracted or not. Just browse it and you can get the information. So, now, in the browse since we extracted 2 variables that is FSU and Segment, so data with two variables have been extracted now. Now question arises here how to define a dictionary file, how to develop a dictionary file.

So, for that you do one thing. You simply write down all those things. In fixed dictionary on a do file you create a do file simply you click here create new and simply clear.

(Refer Slide Time: 43:42)



Let us open, since do file has already been opened, it is here you open another do file. And you just write down this way; infix like you simply on this screen infix dictionary and Stata is very case sensitive you have to write very correctly. Then on the next row the bracket has to be started.

Then on the next row all the variables and the another bracket has to be started. Then your variables names though variable name which is displayed on your screen like str you have to write down str then FSU. Where you are going to get this? You are going to get this from the layout file.

The layout file has given all those information like suppose you are extracting the individual file these things has to be written. Now first of all you should get the minimum information like these in the README file I already said you that your common primary information has

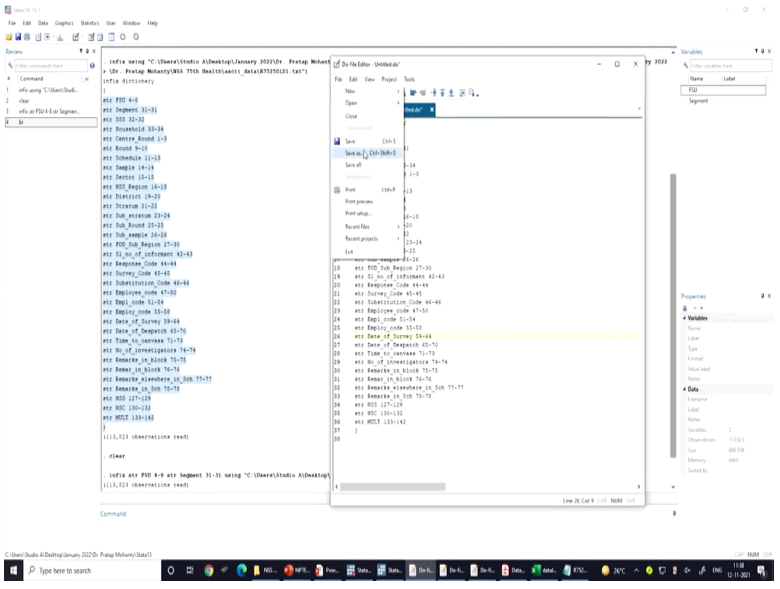
to be given FSU, Segment number, SSS number, household number, level number everything you should keep it first.

Keep it first, its byte positions are given 4 till 8 that is 4, 5, 6, 7, 8. Till 8 position you need to write down 4 to 8 on the first screen it is here. So, like str we have defined str FSU 4 to 8 then str Segment 31 to 31 and etc. Your all those variables will follow at the end you should remember that in a multiplier file should also be written very correctly.

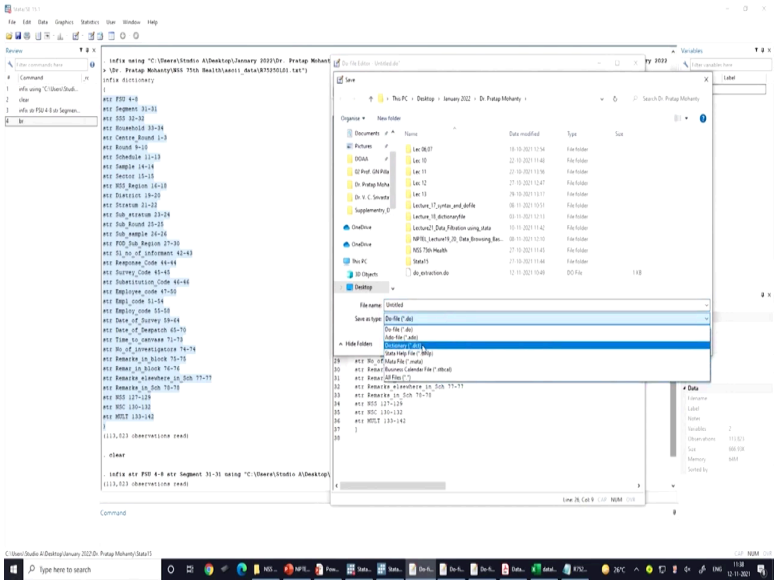
Multiplier files has already been given in each entry serial number 24 here, on level number 3 multiplier 5 position number 132, to 142. So, that has to be also defined very clearly. So, this is 132, 142 is defined and at the end the bracket must be closed. If we can just copy it and paste it over here and, that is all about our dictionary file.

And you can manually do initially then later on that is what is our dictionary file. So, bracket has to be similar in both the case. The starting bracket and the end bracket has to be same. So, it has read that you have given the entries correctly.

(Refer Slide Time: 46:34)

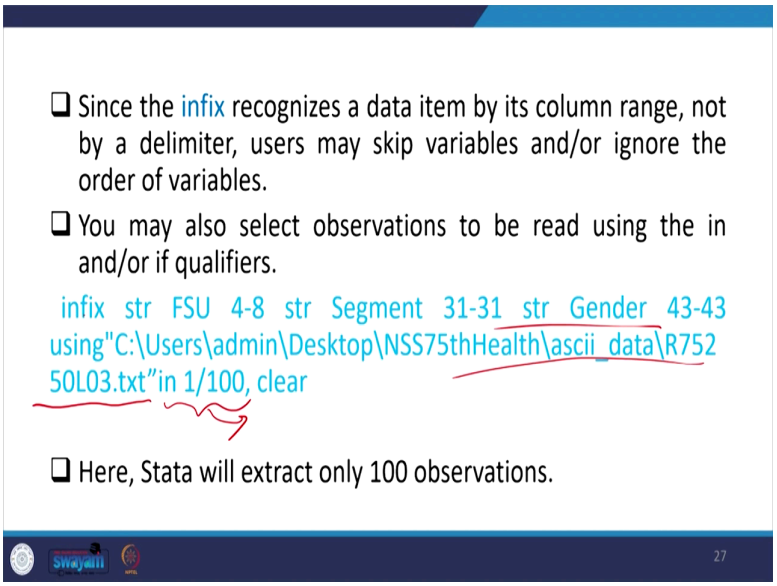


(Refer Slide Time: 46:35)



How to save this? You need to go to Save As, but on the Save As, by default it takes do file you need to change it to dot dct and that if you are extracting the information about household then you write down household dct or level 3, level 2 dct. You have to write down the appropriate name here. Then you save it and at the time of the appropriate command you can give the command and do the extraction. So, rest it is easy, I have already guided you.

(Refer Slide Time: 47:05)



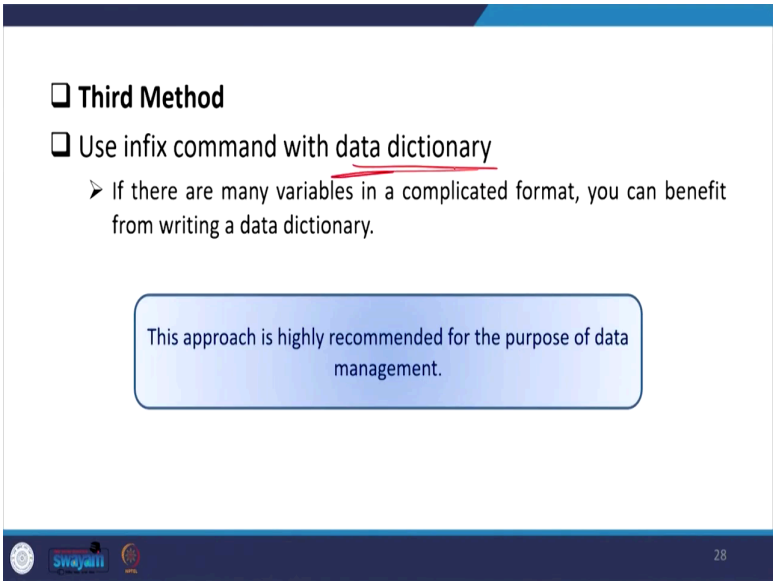
So, these are all I think I have said Stata will extract. In this case suppose another aspect like if you do not want entire information, you want to get very specific information from 1 till

100 observations then it has to be mentioned. Like infix string FSU it is everything you have defined.

And the string position number you have already defined. At the end if you want that you need to extract 1 upon 100 or starting till 100 observations if you have to specify this then your data will be extract till 1 to 100.

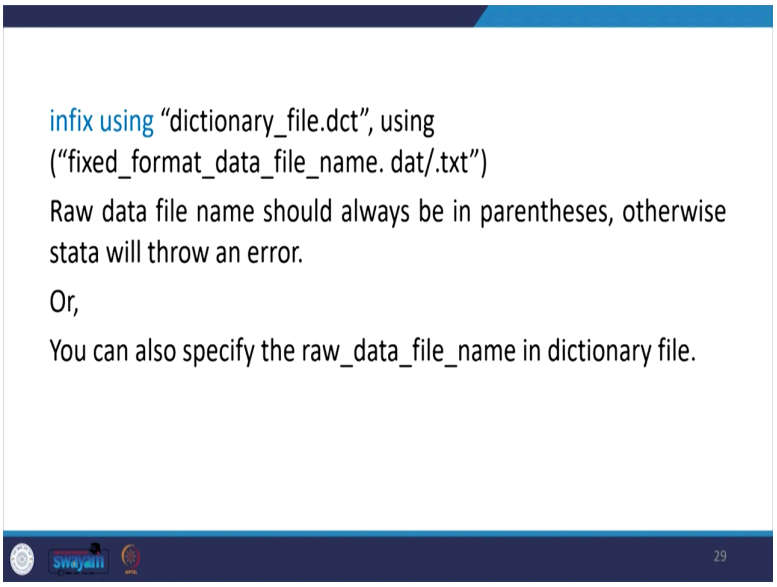
Generally, we do not suggest this because we always suggest that you, please download extract all the information from the beginning. When you are doubly sure about your data and your data is confined to 100 observations only then you can go for this approach.

(Refer Slide Time: 48:09)

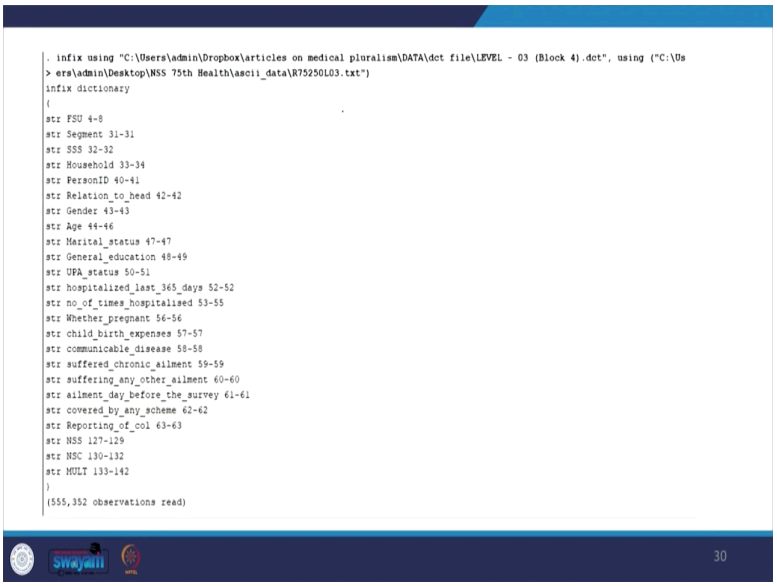


Third method is highly recommended that is the dictionary method. And though initially it seems to be complicated, but actually it helps us in getting a better direction and bigger databases can be extracted in a every less time.

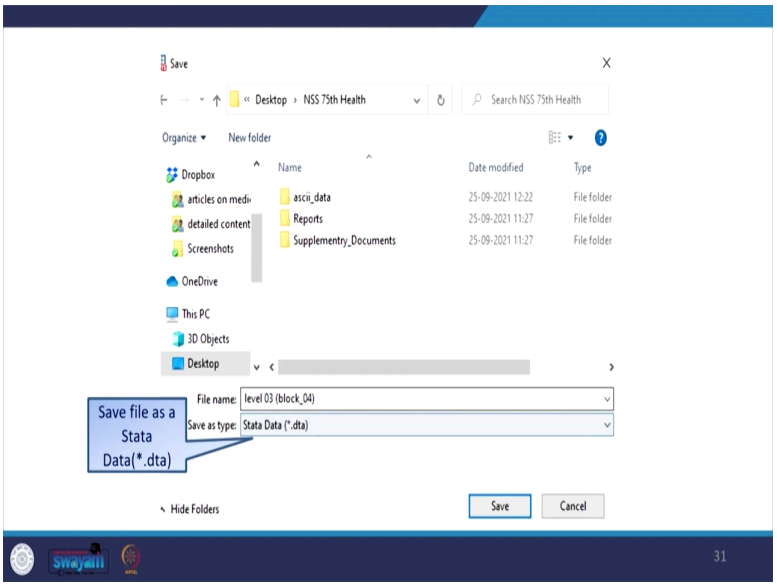
(Refer Slide Time: 48:27)



(Refer Slide Time: 48:43)

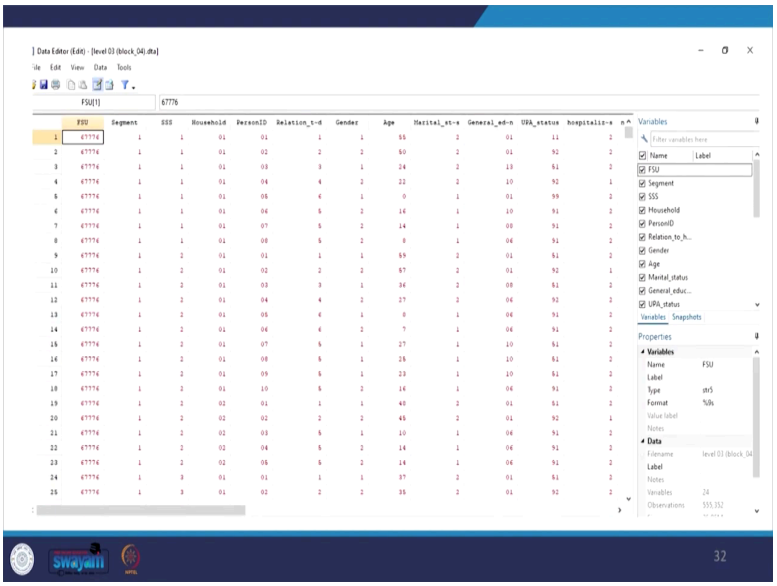


(Refer Slide Time: 48:48)



So, dictionary file the way we have already defined and saved it I think it will be very useful to you. So, infix using a dictionary file is required once you have created your dictionary file and the path name of the data should be also given in the command. This is how it looks like, and we have shown it just a couple of minutes back and then you need to save it accordingly and it will be saved.

(Refer Slide Time: 48:56)



So, this is again repetition of my guidance. I have shown how these variables look like and how you can go for extraction alright. So, if you have any difficulties you may redirect to us,

or you may give all your queries in the chat box of your NPTEL path. And we will be very happy to address it and with this format you can able to extract all the database of NSS and there will be no difficulties.

But I suggest that you should do enough practice the way we have guided and rest all those supporting files with it is PPT with commands layouts or everything we have guided and you should make everything ready before extraction. So, these are all for today and on the next class we will be coming up with content for systematic guidance for you to merge the data sets. And then the next class onwards we will be ready for analysis of the data.

Thank you.