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

Lecture - 18
Making Dictionary File

Welcome friends once again to my NPTEL MOOC module on handling Healthcare data. We are on the 4th week. Myself Dr. Pratap Mohanty; I have been teaching in IIT Roorkee over the last 5 years. I have had experience of more than 14 years in teaching and research. This module has been generated over my experience of teaching in research methodology courses and various data programs.

This module is assisted by our team members Mr. Milind and Mr. Kamal, and they will also be helping you to clarify the doubts. So, let us understand further practical approach. Let us understand the concept or operation STATA or get the practical understanding of Stata data handling.

So, now on this particular lecture, we have kept the title as Making Dictionary File, how we can prepare a dictionary file. This is I think I have guided in my previous module as well; but since this is new in this particular module, I will guide you very carefully and how you guys can be able to go for it. So, what do you mean by dictionary file?

This is basically a template or a set of instruction that tells software how to read your data. Why do I require a dictionary file? That might be an obvious question comes to your mind. Like you might be going through the raw data of a national sample survey; national sample survey whenever releases data, they provide data in ASCII format or the raw data.

After some years they will provide the extracted data in STATA format or even a SPSS format, in SAS format; but at the starting point they provide you free access of the raw data. When the raw data is in front of you, we will be guiding you how you should go for extracting the raw data. To extract the raw data, there are different options available; we will also guide you correctly, how to extract the raw data and how it looks like.

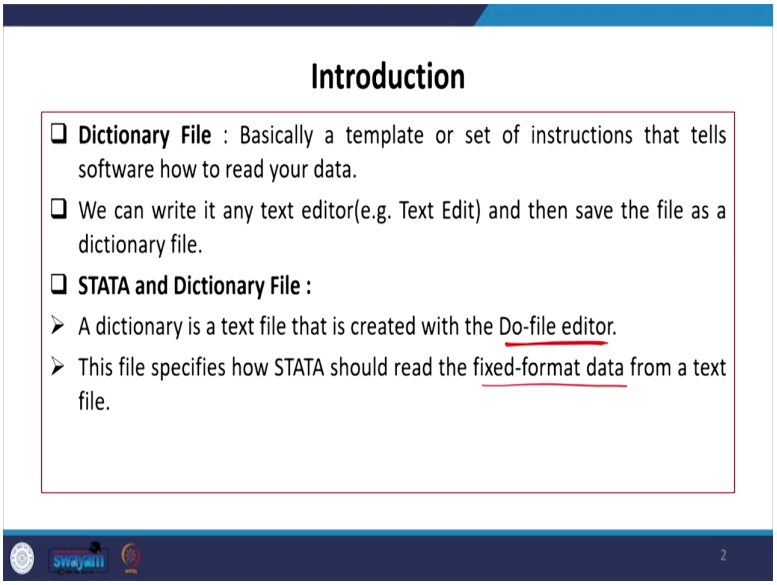
I think I will be guiding you in detail, but at this moment I am giving you a very basics of it. When your raw data is in is following a format, you will actually follow the appropriate

extraction technique. Format calculated like delimited format, free format and fixed format; there are different data, we will explain all those things later, but not in this lecture.

When those data are available in different format, you require a different type of extraction. In the extraction process, you would certainly require a dictionary file; dictionary file means it sets or keeps all the commands in a systematic format, so that your extraction would be very easier and there would not be any possibility of mistakes.

That does not mean without dictionary file, you cannot able to extract; that it is wrong guidance, you can also able to do it, but dictionary file helps to make your process faster. So, that is why I said that the dictionary file is in fact a template or set of instruction that tell Stata software how to read your data.

(Refer Slide Time: 04:44)



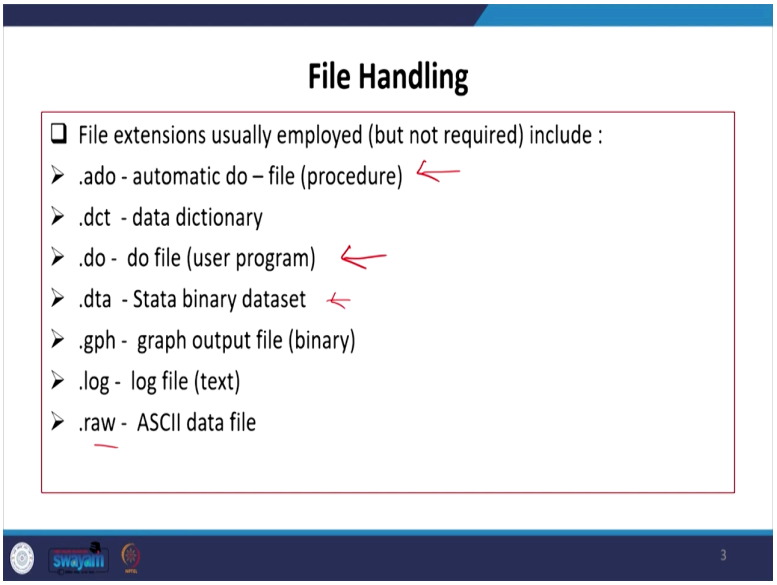
We can write it any text editor, I mean in any text editor or in the do file that looks like, that is in fact a text editor or text version.

And, then if you save it, save the file as a dictionary file; at the time of saving, if you save it with dot dct or dictionary file, you are actually making a dictionary file finally. Then, second aspect is like what about STATA, dictional file in STATA; a dictionary file is a text file that is created with the Do-file editor.

So, when you open through the Do-file editor in STATA, that is basically called a dictionary file in STATA. This file specifies how STATA should read the fixed format data from a text

file. So, we will be explaining you about this later, but fixed format basically the byte position of the variables and its exact location is actually well defined in the dictionary file.

(Refer Slide Time: 06:00)

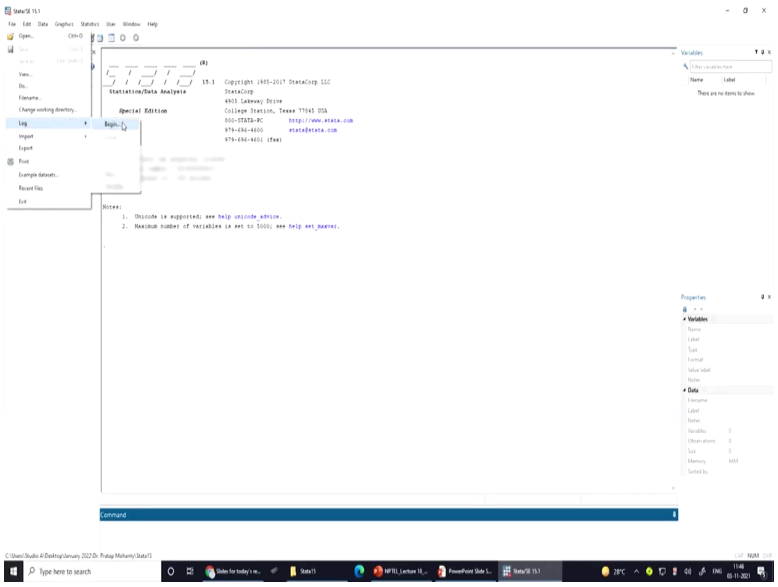


Like in file handling there are different file handling required in in the process. There are different file handling like file extension, usually employed, but not required enough; but like a dot ado file, ado file is all about automatic do file. So, far we discuss about do file, where we manually entered our command that simply called dot ado, dot do file, that is basically user-based program.

But if it is automatically designed and developed procedures, they are called ado file. Dot dct, it is the dictionary file, which we are going to tell you and guide you; then dot dta basically is all your STATA binary data set; if it is in binary data set, then that is called dot dta, then dot gph basically graph output file.

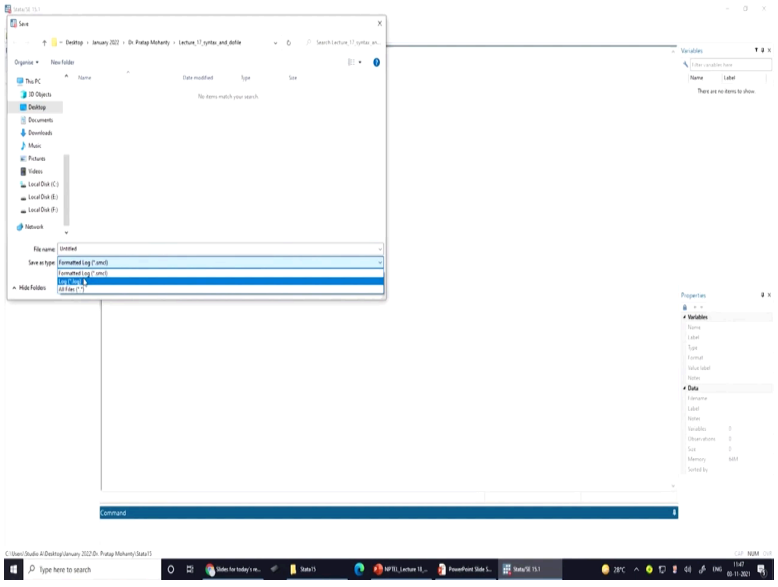
Graph output file if it is there, it will come with dot gph. Then dot log, log file basically I have already shown to you; I will also show some of those details through the STATA. The last one is called dot raw file, basically the ASCII data it, it comes with a extension with dot raw. So, I can just show it to you in STATA window, it is here in our files.

(Refer Slide Time: 07:50)



We are opening STATA here on the screen. Now, I will show you that how you can see different file in the do file. In the do file itself you can say; like first one I will show you the log one, which I have already told you. Go to file, then you can open the log; suppose log begin, at the start I think I told you earlier that, when you start operating STATA, you simply open the log file, every time.

(Refer Slide Time: 08:15)

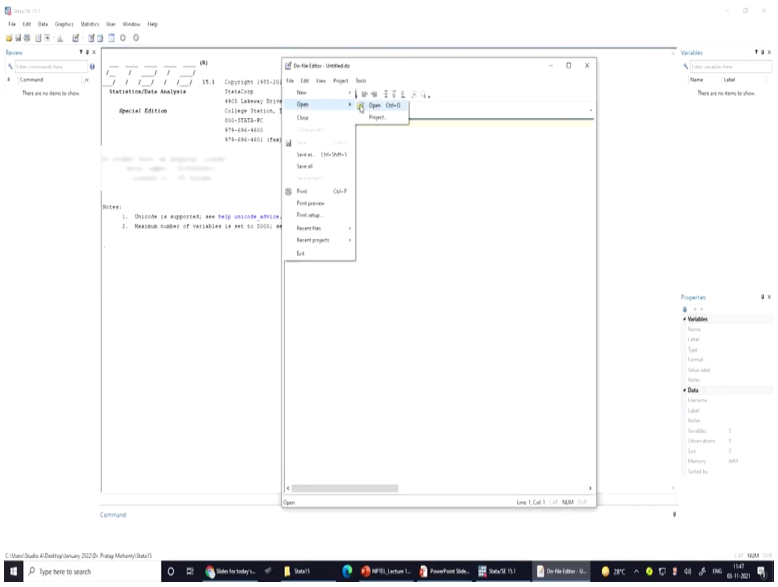


And you do save with your file name here; it will come with the extension like dot smcl. Smcl that does indicate your STATA actually whatever the operation you do on the screen,

the same thing will be actually stored with dot smcl; anytime you open your screen, it comes with the same results. Some results with same look, likewise your STATA window gives.

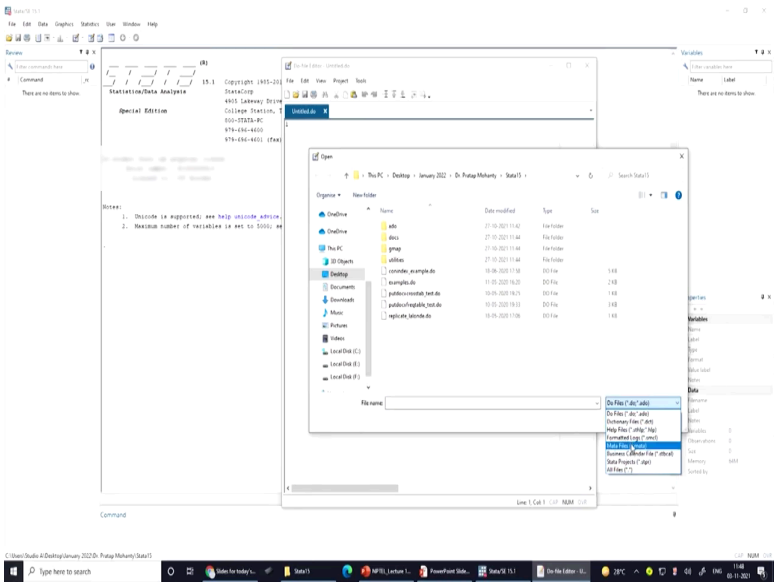
But if you save in dot log, it is like dot log; dot log it actually saves in text file, dot txt file. So, this is your first guidance.

(Refer Slide Time: 09:12)



And second, we will actually show it through a do file. In the do file, if you just like save it; there are different versions.

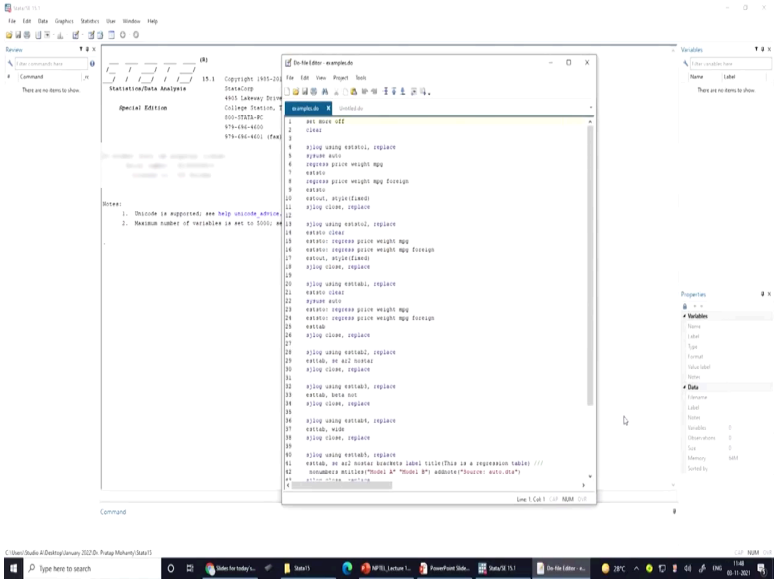
(Refer Slide Time: 09:36)



Like, suppose I open one do file on your screen. Like here on the screen I am just going to tell you; you just see what other files are available. On the screen like dot do or dot ado, then dot dct as the dictionary file, then if it is dot smcl file.

Then, there are many other directions; we have not gone through, I am not guiding much. But at this suppose I just open one do file on the screen may be this one, maybe just this one for your clarity.

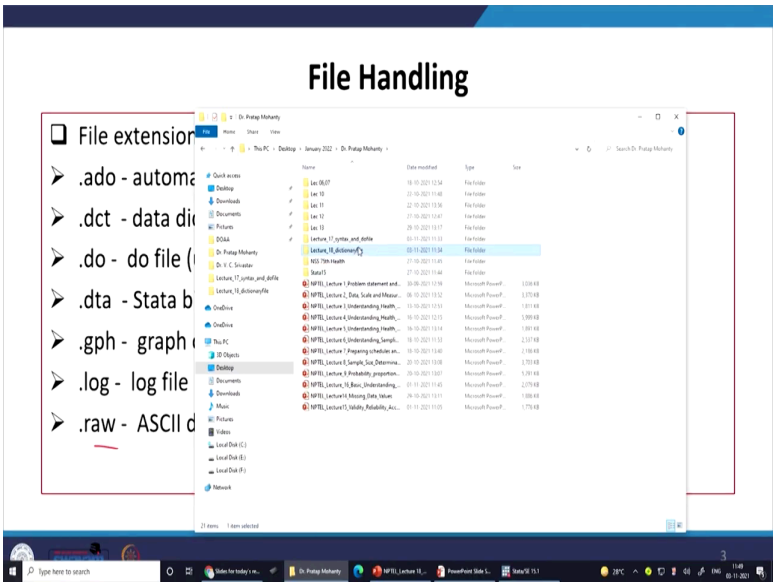
(Refer Slide Time: 10:12)



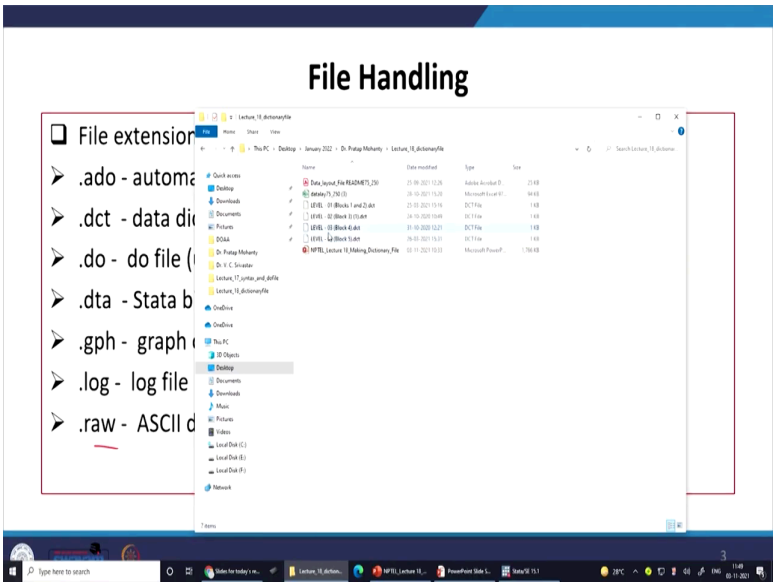
There are so many commands. You can save it and change this to other version if it is suitable to you; I will clarify rest of the detail later, but at this moment we are not guiding.

So, let me just move it to our required guidance once again, here it is ASCII file I think I have not guided, let me just show you where it is located in our data, it is here, let me go to my data once again.

(Refer Slide Time: 10:48)

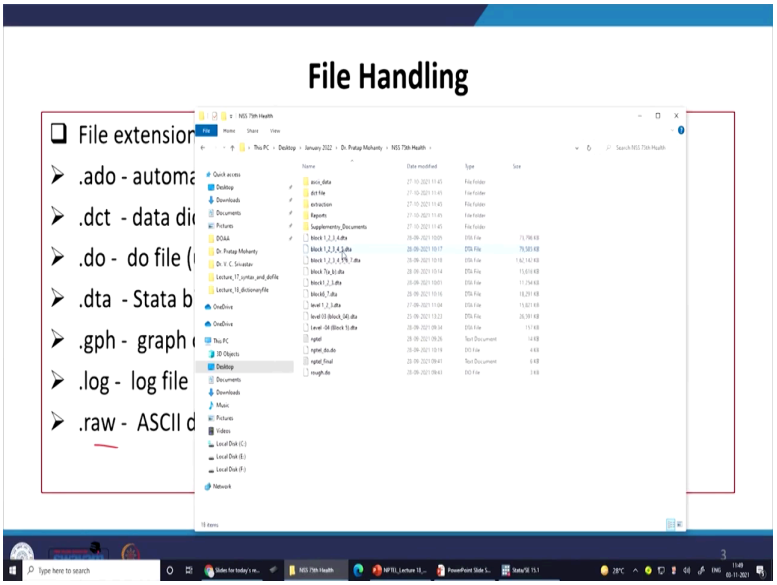


(Refer Slide Time: 10:50)



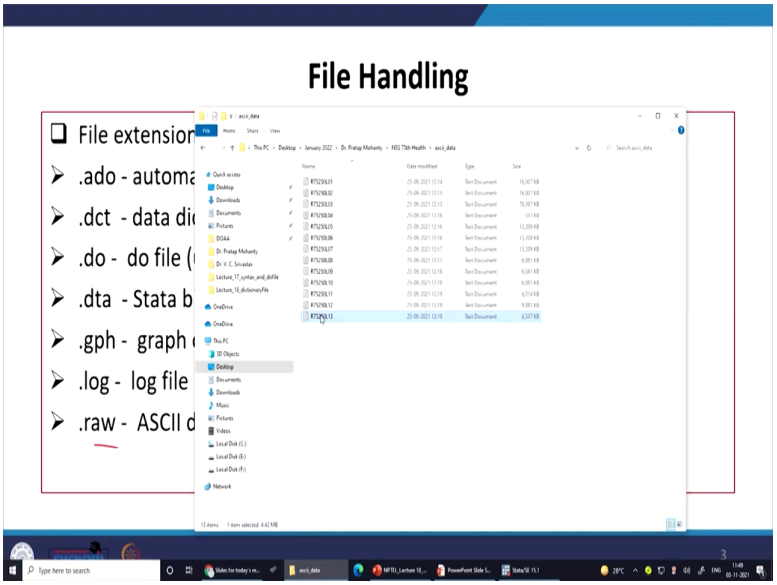
It is the lecture number 18 we are trying to find out. Now, the dot dct file we have already made and I will also clarify how to make it, which way we should go for it.

(Refer Slide Time: 11:12)



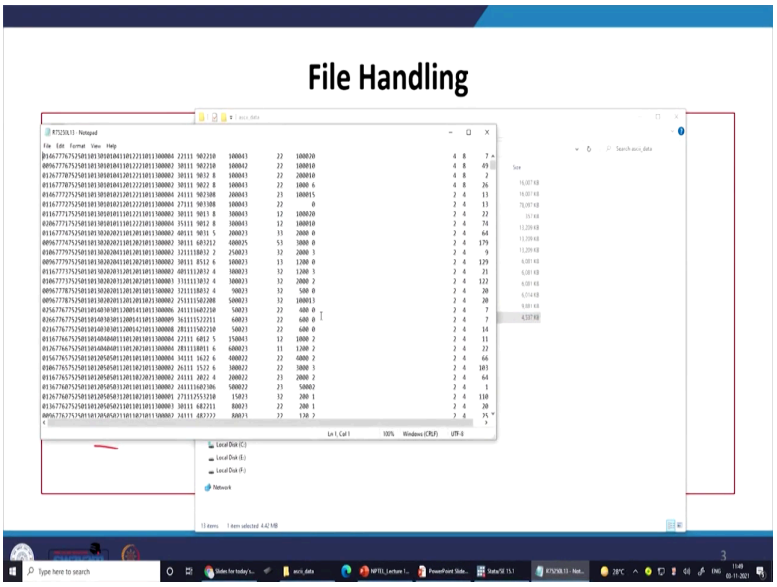
And like in my data I have already shown to you, that what is about your raw data, ASCII data.

(Refer Slide Time: 11:16)



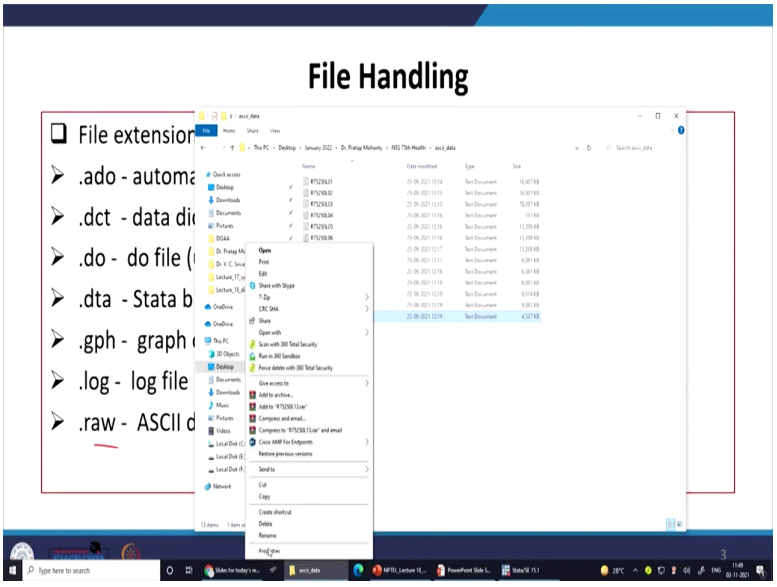
Here is your ASCII data. Now, just right click any one file.

(Refer Slide Time: 11:27)



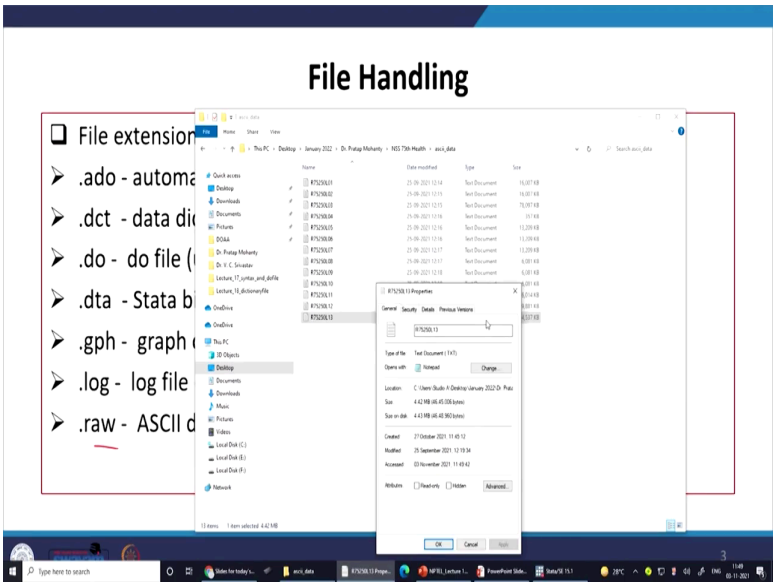
Simply close it and right click on it.

(Refer Slide Time: 11:33)



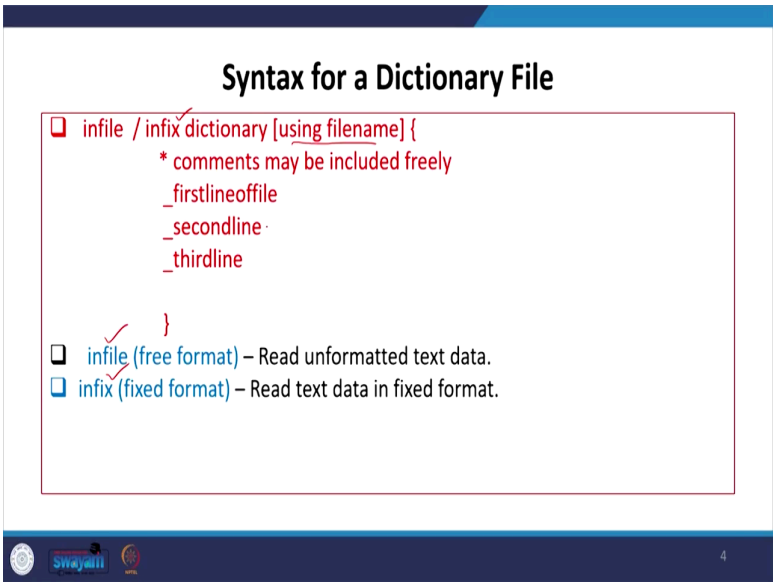
Go to the properties at the end.

(Refer Slide Time: 11:36)



Now, it tells you what is the format, the format here is dot txt. So, raw data you can find sometimes it comes with dot raw, some format they give it in dot raw or dot txt.

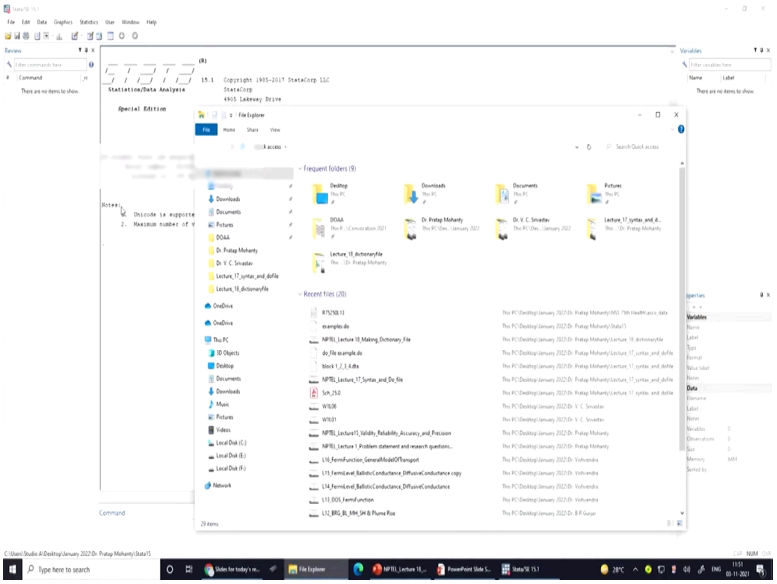
(Refer Slide Time: 12:01)



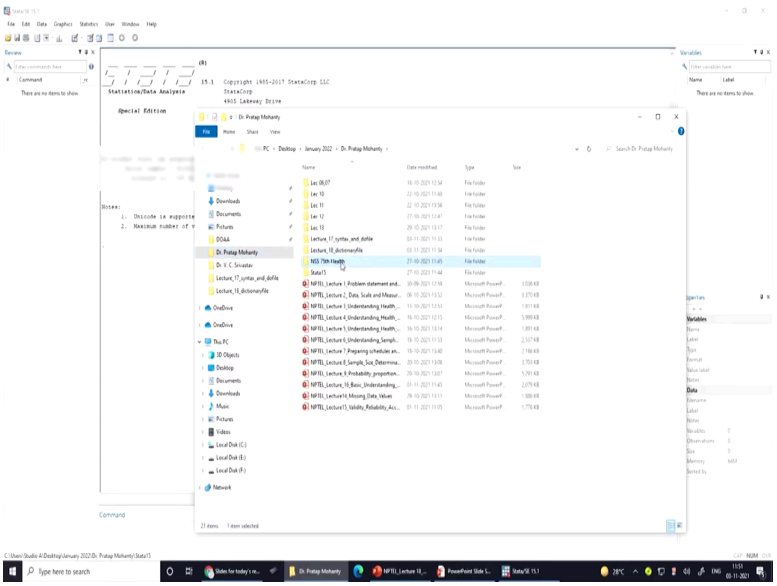
Some other details I have already guided, so no need to spend much time on it. Now, we are going to make a dictionary file on the screen. So, what does it indicate? It indicates whether your dictionary file is of infile format or infix format.

In fixed format means, it is a fix format data set, I will also show it; just once again I am opening it here, suppose I just open the data for you.

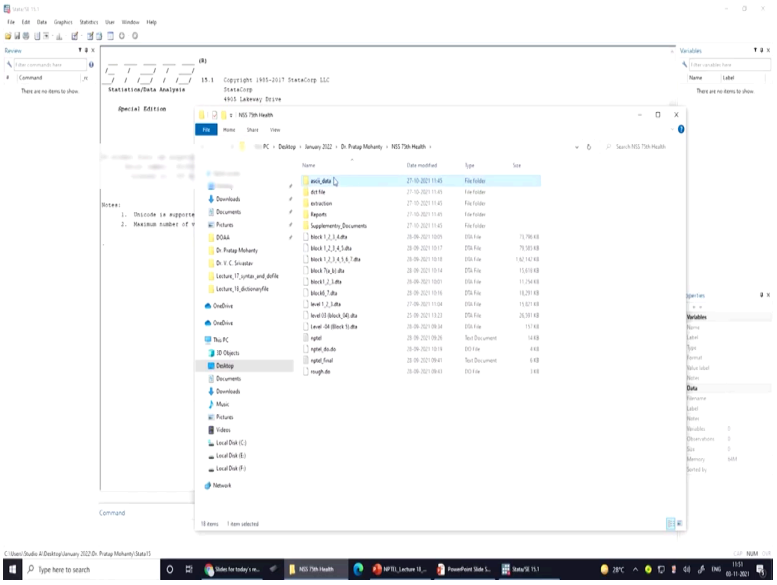
(Refer Slide Time: 12:35)



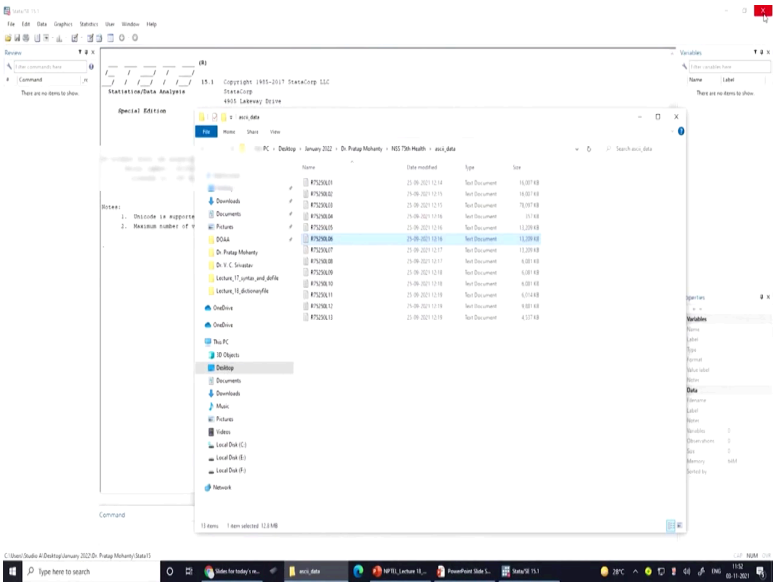
(Refer Slide Time: 12:39)



(Refer Slide Time: 12:41)

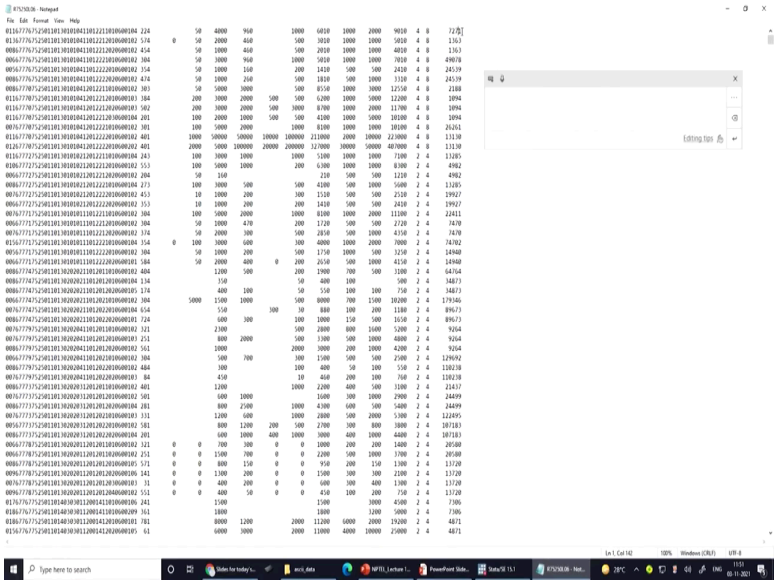


(Refer Slide Time: 12:43)



Now for your understanding, I have opened one raw data.

(Refer Slide Time: 12:47)



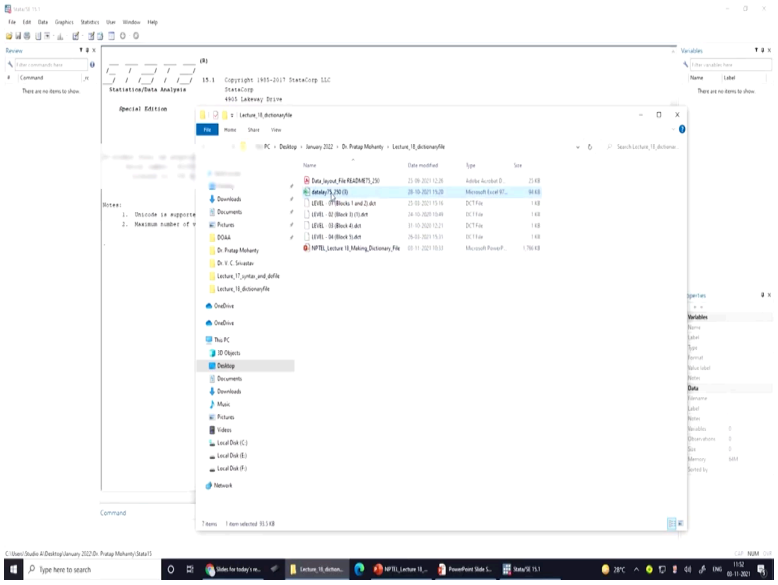
Just have a look on the raw data; there are different entries, starting from first entry till end entry.

Now, if I just click here, you can read what is all about these entries. At the end if I just take it to the last point. Now, at this one you can read on the bottom of this page like column number 142, line number 1; anywhere suppose I enter, I keep my cursor here and I have entered here, now this suggest your line number 44 and column number 142.

Similarly, here, it is saying column number 38. Why am I explaining this? Because some variables are defined with their column position. Column position and there are some blank space given deliberately to identify the variables and their entries. May be in some case there are no responses; there are missing responses, but missing responses are also equally useful for research.

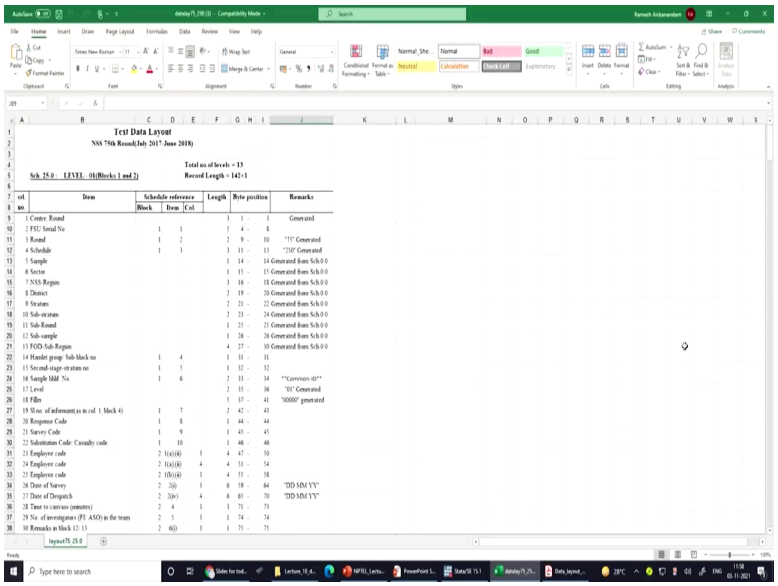
But it is clearly indicating all position very correctly. How to know that what are those positions? We have to go to your layout file.

(Refer Slide Time: 14:15)



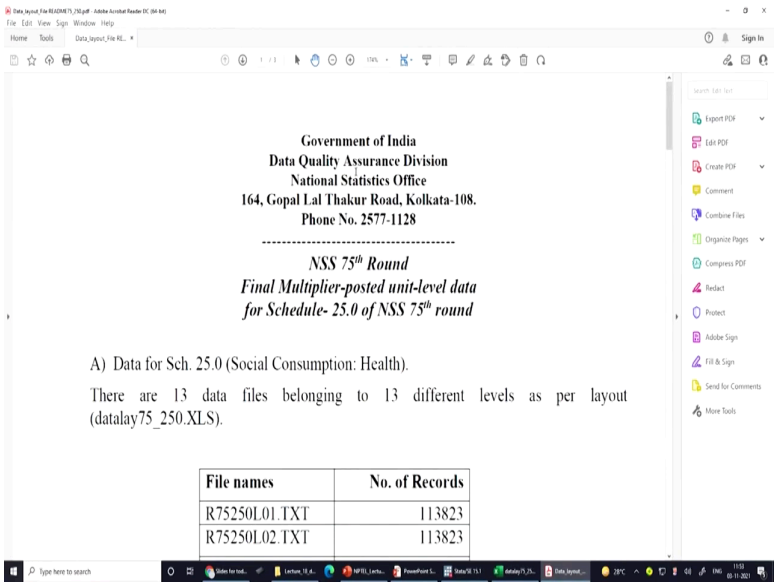
In the 75th round, you have got the layout file; we can also open the layout file of that 75th round.

(Refer Slide Time: 14:21)



So, now here it is, here is your layout file on the excel page; now also I am simultaneously going to open the data layout readme file.

(Refer Slide Time: 14:33)



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

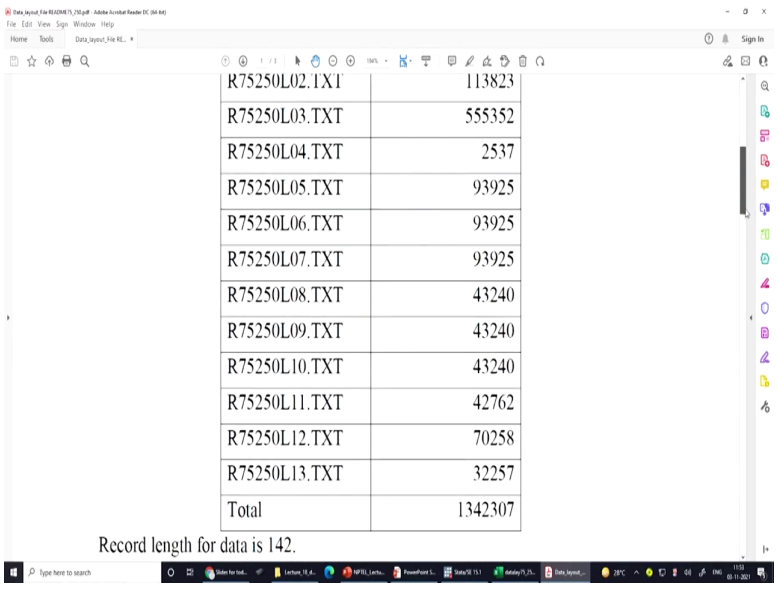
A) Data for Sch. 25.0 (Social Consumption: Health).

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

So, readme file how it is important in STATA and specially for NSS 75th round.

(Refer Slide Time: 14:41)



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.

This is very important and equally challenging for the newcomers or the new researchers, who wanted to do some study. The NSS data provides you the readme file, the ASCII data or that is the raw data, all those raw detail in different blocks.

It also provides you the multiplier file, it provides you the guidance about the codes, different name of the state and codes, etc. Now, on the readme file of the 75th round specially, this is

very important. We are again highlighting that schedule number 25 of NSS 75th round. Now, this is on social consumption on health care; there are 17 different levels provided by STATA.

Why are levels defined? Levels are indicating different special information; but from the initial entries it gives you level number 2 is your household, household characteristics. Household characteristics what does this mean? This means it has asked the questions about how many members are staying in the household, what is your religion, what is your caste.

All such household characteristics are asked in this particular questionnaire or in this set of questions. On the 3rd level it is called individual characteristics or the characteristics of a household member and his or her responses about the health care. These clarifications are already given in the report, you have to read the report or the questionnaire or the schedule.

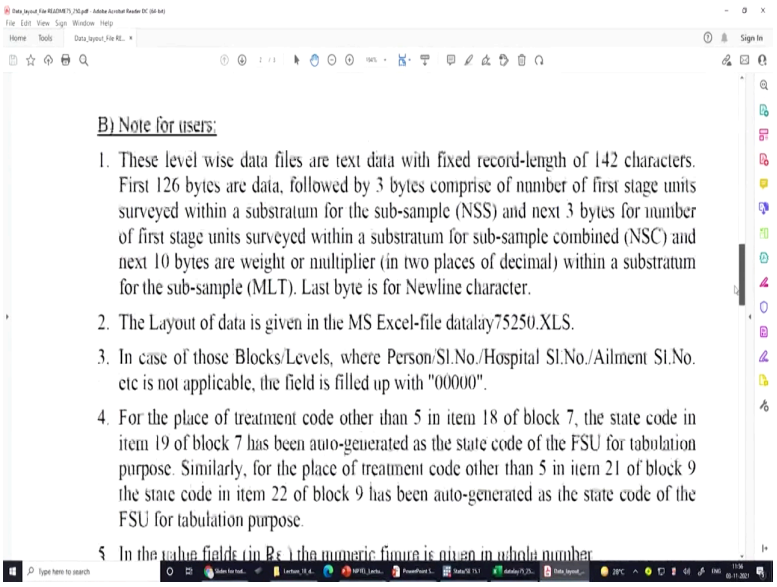
So, you also have to open your schedule simultaneously while reading between the lines of these entries. Now, you might be surprised that why it is much higher in individual i.e., 555352, whereas in the household it is lesser; it is quite obvious that within the household there might be so many members.

Since all the members are covered, there are more numbers expected; therefore, the members and their characteristics are more by entry. But on the 4th level the entry is too less. Why it is too less? Because the 4th level is indicating information about death, i.e., death of a person in last 365 days. So, in last 365 days the questions were asked: Is there any members who passed away during the last 365 days?

So, obviously the information related to deaths would be lesser; that is why the number of entries are in fact very less. Similarly, there are short term morbidity, long term morbidities; like for 15 days elements, 365 days elements, so many things are provided in this round. Where to get the layout? Layout you can download over here.

There is a hand-mark here, if you click it, you will get the layout; if you click on those things, you will be redirected to another page to download this txt file.

(Refer Slide Time: 18:23)

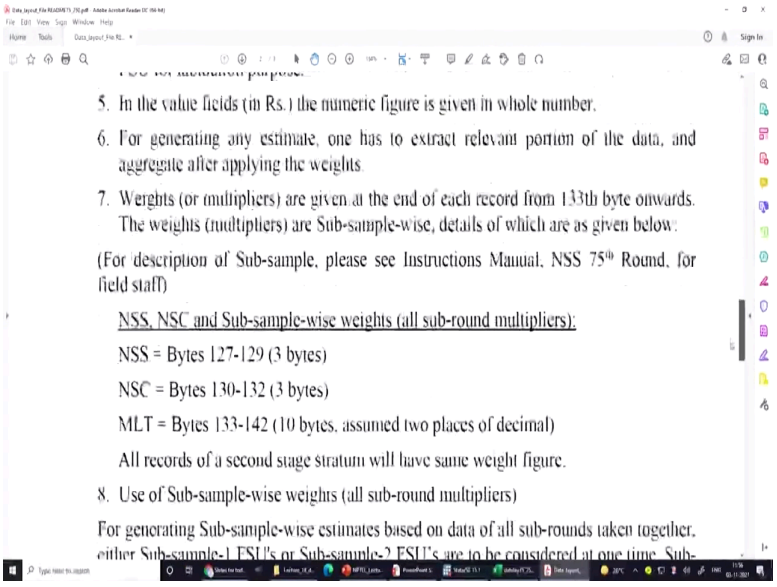


B) Note for users:

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.

So, all those things I have guided.

(Refer Slide Time: 18:24)



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.
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)

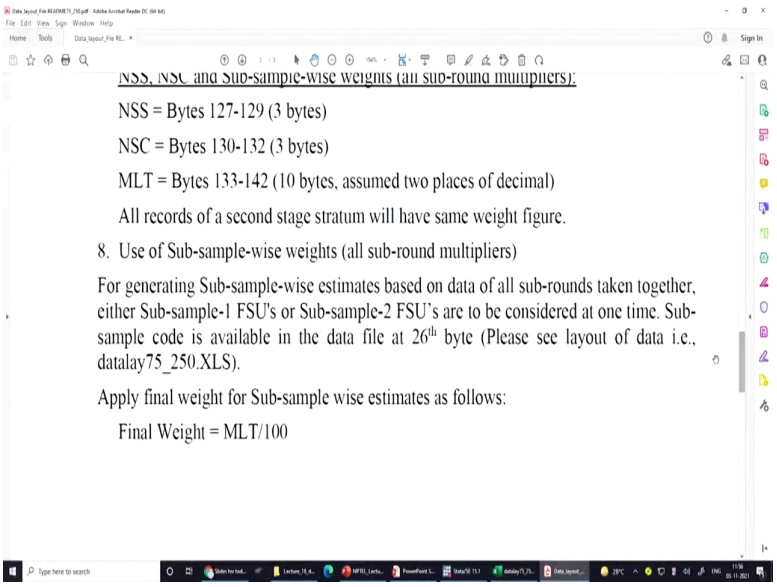
MLT = Bytes 133-142 (10 bytes, assumed two places of decimal)

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-

(Refer Slide Time: 18:25)



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)

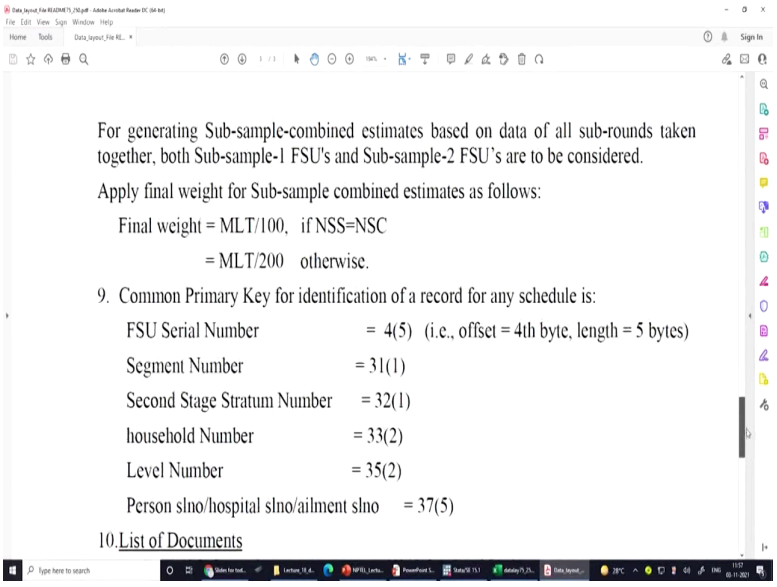
For generating Sub-sample-wise estimates based on data of all sub-rounds taken together, either Sub-sample-1 FSU's or Sub-sample-2 FSU's are to be considered at one time. Sub-sample code is available in the data file at 26th byte (Please see layout of data i.e., datalay75_250.XLS).

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

Final Weight = MLT/100

This readme file also gives information about your multipliers.

(Refer Slide Time: 18:29)



For generating Sub-sample-combined estimates based on data of all sub-rounds taken together, both Sub-sample-1 FSU's and Sub-sample-2 FSU's are to be considered.

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

Final weight = MLT/100, if NSS=NSC

= MLT/200 otherwise.

9. Common Primary Key for identification of a record for any schedule is:

FSU Serial Number = 4(5) (i.e., offset = 4th byte, length = 5 bytes)

Segment Number = 31(1)

Second Stage Stratum Number = 32(1)

household Number = 33(2)

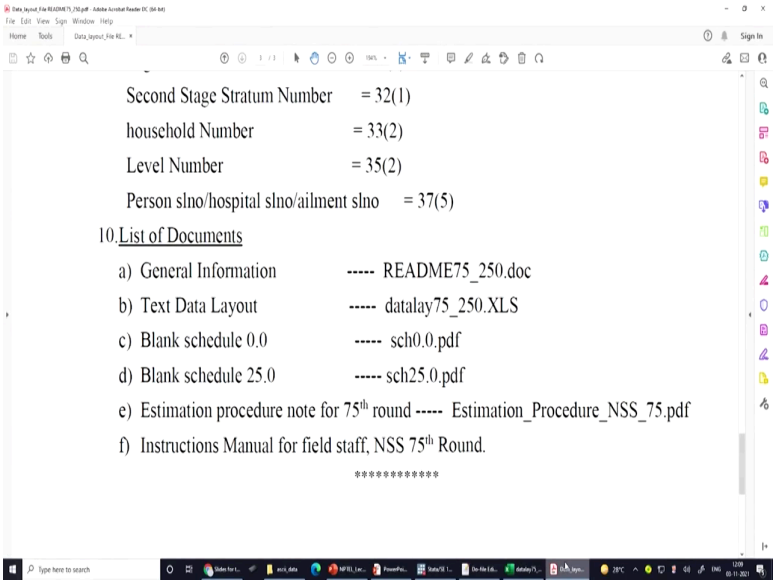
Level Number = 35(2)

Person slno/hospital slno/ailment slno = 37(5)

10. List of Documents

How your multiplayer looks like?

(Refer Slide Time: 18:30)

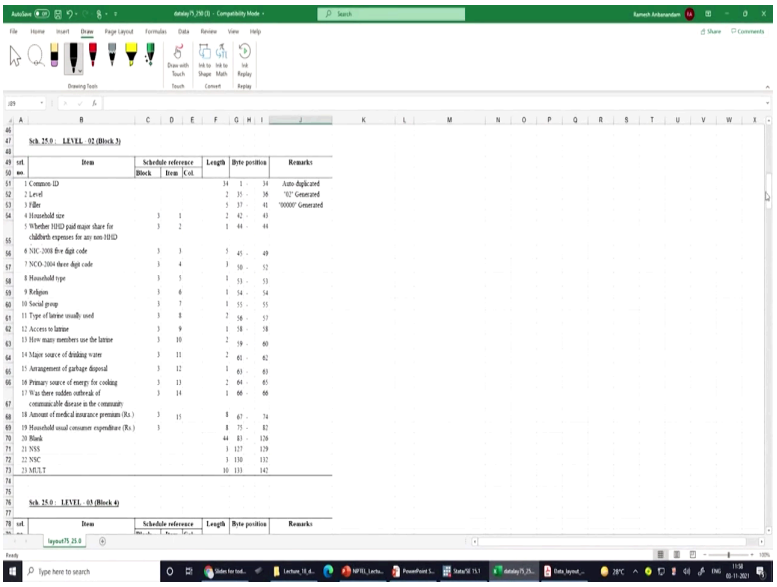


Then at the end, I will be emphasizing for sure about defining a common primary id. Since we have different layers, different levels, how each level can be merged together. Why merging is required? Because as a researcher we cannot able to just do our research based on the household characteristics or any individual file. We have to merge another file with the first one or second with the third, third with the fourth or we can combine all the different blocks of information.

So, this is essential and this primary key you must remember very carefully; this byte positions are given 4 till 5th position, like 4 till 8 on the entry 4, 5, 6, 7, 8. So, 5 positions indicating your FSU serial number. Then segment number, second stage stratum number; these clarifications are given in the report, you can follow the report.

Read first ten, fifteen pages of your report, you will get enough idea about it. So, what are very essential for extraction? Readme file, data layout file, schedule file and report is quite essential. So, now, I am going to explain you about the data layout, before discussing with you dct file or dictionary file. Data layout as I already told you it comes with levels, level 1 its information are given.

(Refer Slide Time: 20:11)



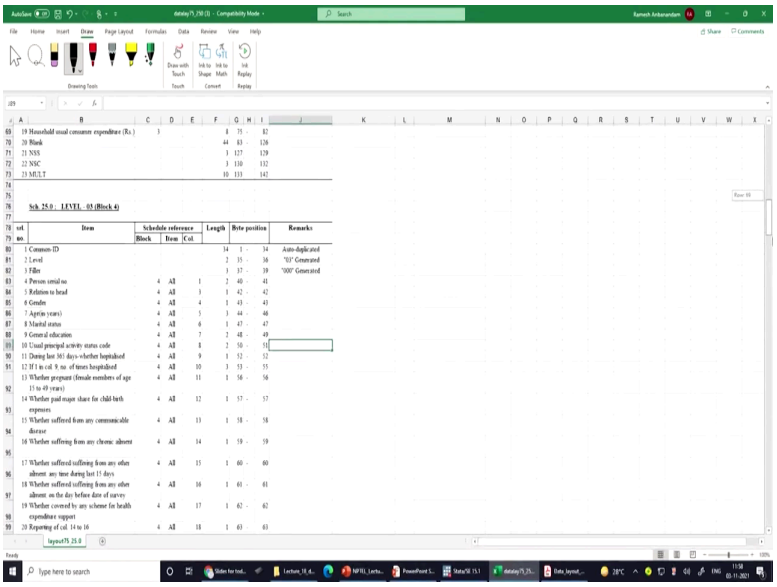
Sch. 25 0 : LEVEL - 02 (Block 3)

srl. no.	Item	Schedule reference Block	Item	Col	Length	Byte position		Remarks
1	Common-ID				14	1	14	Auto-duplicated
2	Level				2	15	16	"02" Generated
3	Filler				5	17	21	"00000" Generated
4	Household size	3	1		2	42	43	
5	Whether HHD paid major share for childbirth expenses for any non-HHD	3	2		1	44	44	
6	NIC-2008 five digit code	3	3		5	45	49	
7	NCO-2004 three digit code	3	4		3	50	52	
8	Household type	3	5		1	53	53	
9	Religion	3	6		1	54	54	
10	Social group	3	7		1	55	55	
11	Type of latrine usually used	3	8		2	56	57	
12	Access to latrine	3	9		1	58	58	
13	How many members use the latrine	3	10		2	59	60	
14	Major source of drinking water	3	11		2	61	62	
15	Arrangement of garbage disposal	3	12		1	63	63	
16	Primary source of energy for cooking	3	13		2	64	65	
17	Was there sudden outbreak of communicable disease in the community	3	14		1	66	66	
18	Amount of medical insurance premium (Rs.)	3	15		8	67	74	
19	Household usual consumer expenditure (Rs.)	3			8	75	82	
20	Blank				44	83	126	
21	NSS				3	127	129	
22	NSC				3	130	132	
23	MULT				10	133	142	

Sch. 25 0 : LEVEL - 03 (Block 4)

srl. no.	Item	Schedule reference Block	Item	Col	Length	Byte position		Remarks

(Refer Slide Time: 20:12)

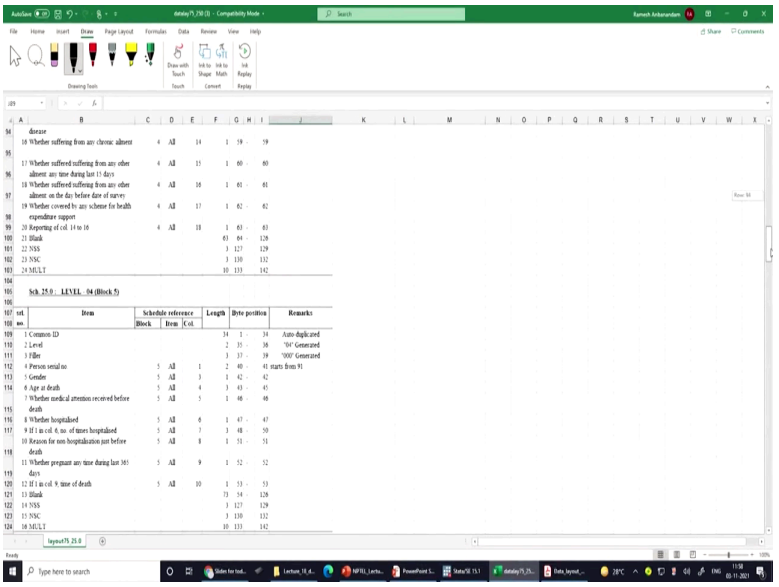


Sch. 25 0 : LEVEL - 03 (Block 4)

srl. no.	Item	Schedule reference Block	Item	Col	Length	Byte position		Remarks
1	Common-ID				14	1	14	Auto-duplicated
2	Level				2	15	16	"03" Generated
3	Filler				3	17	19	"000" Generated
4	Person serial no	4	All	1	2	40	41	
5	Relation to head	4	All	3	1	42	42	
6	Gender	4	All	4	1	43	43	
7	Age(in years)	4	All	5	3	44	46	
8	Marital status	4	All	6	1	47	47	
9	General education	4	All	7	2	48	49	
10	Usual principal activity status code	4	All	8	2	50	51	
11	During last 365 days-whether hospitalised	4	All	9	1	52	52	
12	If 1 in col. 9, no. of times hospitalised	4	All	10	3	53	55	
13	Whether pregnant (female members of age 15 to 49 years)	4	All	11	1	56	56	
14	Whether paid major share for child birth expenses	4	All	12	1	57	57	
15	Whether suffered from any communicable disease	4	All	13	1	58	58	
16	Whether suffering from any chronic ailment	4	All	14	1	59	59	
17	Whether suffered/suffering from any other ailment any time during last 15 days	4	All	15	1	60	60	
18	Whether suffered/suffering from any other ailment on the day before date of survey	4	All	16	1	61	61	
19	Whether covered by any scheme for health expenditure support	4	All	17	1	62	62	
20	Reporting of col. 14 to 16	4	All	18	1	63	63	

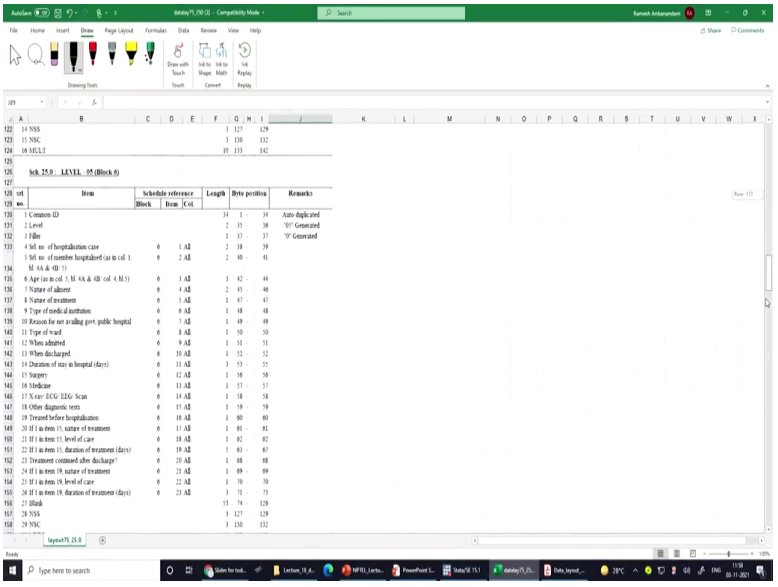
Then here it is your level 2.

(Refer Slide Time: 20:16)



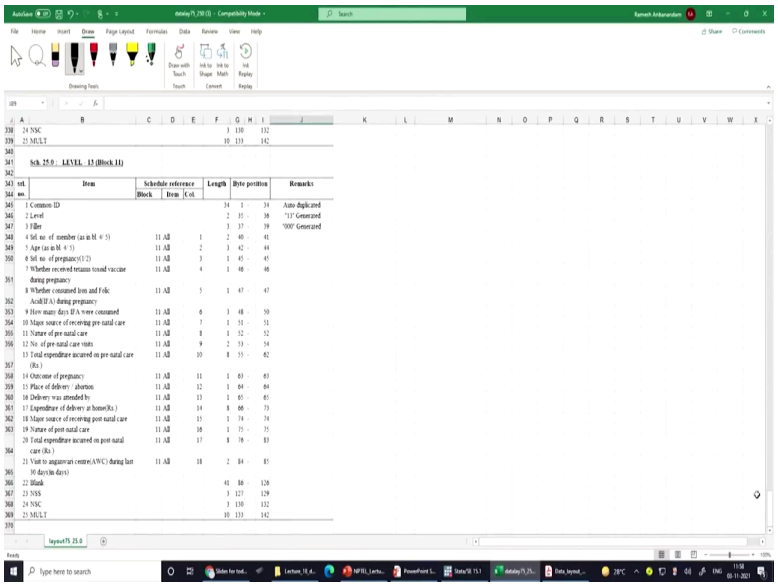
Here then it is level 3.

(Refer Slide Time: 20:17)



Level 4, level 5 and so on till level 13.

(Refer Slide Time: 20:22)



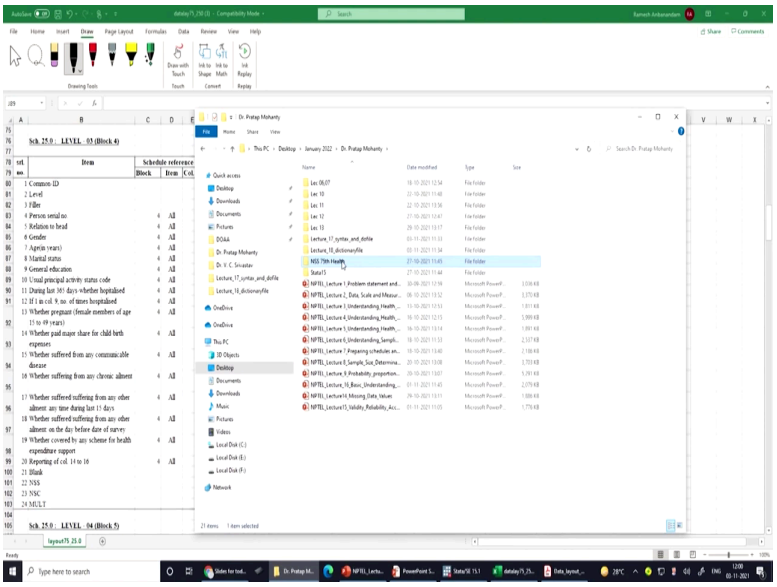
I told you that there are in total level 13, so level 13 is mentioned over here. Then, we can go back to the first one; first one actually carries with all such backgrounds regarding your primary keys and the surveys, how survey was conducted.

So, first sampling unit, then there are how many rounds are there, and what is the schedule number is entered, what are the sectors, how sector rural or urban is entered, those information are in fact very essential. Now, next you quickly should follow your level number; level number 2 indicates household characteristics. How how do I understand it is a household characteristics?

Because it is given information like household type, the religion, their household size; like whether household has major drinking water, source of drinking water or not etc. Then on the next one you have level 3; level 3 indicates the members characteristics, household member characteristics; they are basically age of the person, marital status, education these are the information about a person.

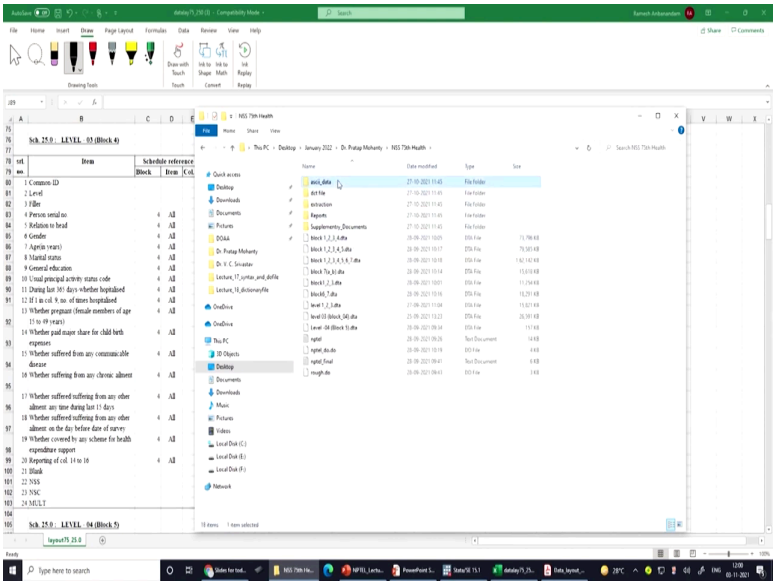
So, what is important after understanding these variables in the file? It is important to look at the byte position, byte position is given like for example, gender; the byte position from the ASCII data it is 43 till 43, 43 to 43 is indicating your gender, gender related information.

(Refer Slide Time: 22:03)

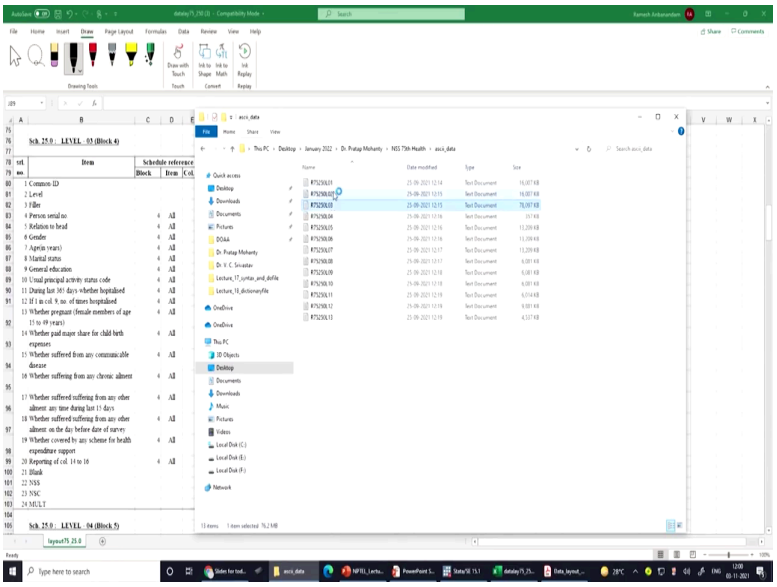


From the data you can easily see, I have already shown to you, it is here.

(Refer Slide Time: 22:16)



(Refer Slide Time: 22:17)



So, now, once again I will go to the 75th round, I will open just a sample information like third for example.

(Refer Slide Time: 22:23)

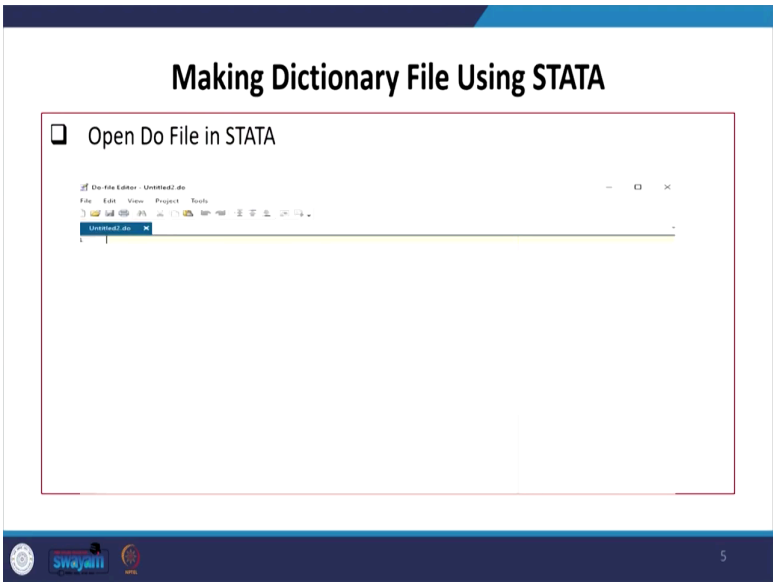
Third file is a big very big file. So, I just said that it is from if I say gender they see; this is in fact 43 till 43, you can just have a check. What is 43 till 43 on the ASCII data, so gender 2 is here, 2 is the entry. So, 2 indicates what? 2 indicates your code for the gender; may be 1 for male and 2 for female. So, similarly on this line 43 to 43, you can check 1, 2, entries are there.

Everywhere you will find 1, 2 etc, so this indicates your gender. Then let us go back to our ppt and I have already clarified all those things and the layout these byte positions are very essential and I am going to use those things for the clarifications. So, why dictionary file is needed?

I told you that based on the data, if it is free format data; free format data, we have to use in file as the command. But since most of our data come with fixed format; fixed format entries, in that case infix command is used. So, nowadays we use infix command; dictionary then using file where you are in fact using these commands.

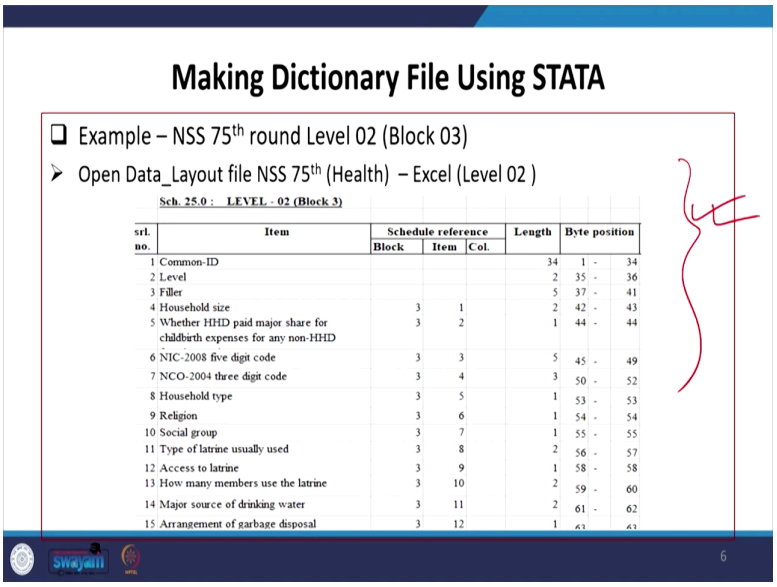
Using file has to be mentioned; then followed by your comments, comments may be included freely by first line and it is file second line and third line I am just going to guide you on the next page.

(Refer Slide Time: 24:29)



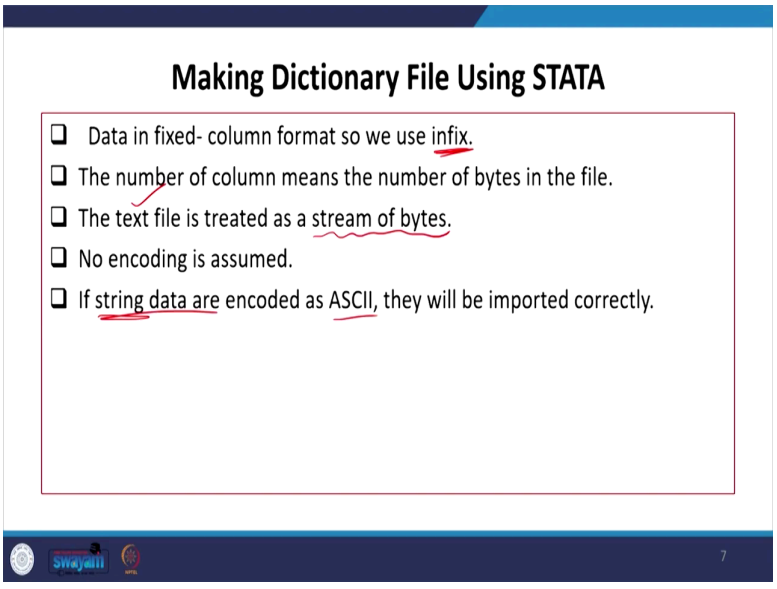
So, first to start with what we will do? We will open the do file in STATA. We will simply open; I am just showing you some of our insert, then I will come back to it.

(Refer Slide Time: 24:41)



So, this is what I have already explained; like your layout file, this is in fact explained very categorically.

(Refer Slide Time: 24:54)



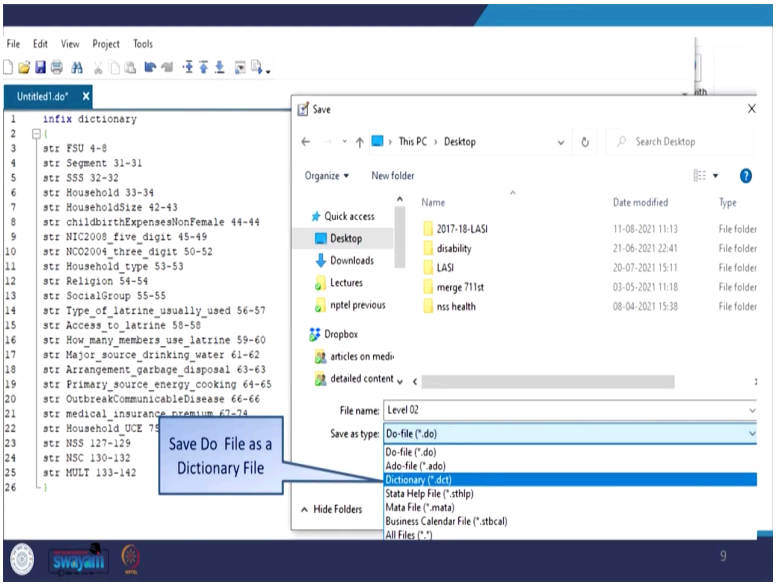
Then how to make the dct file; we will open to do file, I have already opened several times and you can open in between. I am not going to open at this moment, let me clarify first about the do file. Like data if it is in fix format, data in fixed column format, so we use infix command.

I have already told to you; the number of column means, the number of byte position, byte in the file. The text file is treated as a stream of bytes. So, basically the byte space in their stream is indicated in text file, no encoding is assumed. So, that is not assumed; like encoding is in fact is not anything that is given with its clarifications provided.

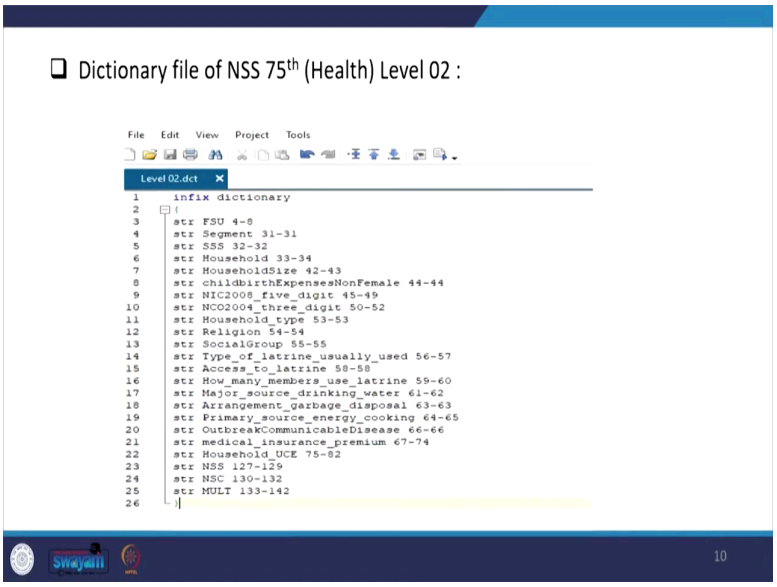
If string data are encoded as ASCII, they will be imported correctly; if it is in ASCII format, then it will be imported. So, string data I will also clarify on the way of explaining what you mean by string data probably from the beginning of my lecture, I have repeatedly mentioned about what is the meaning of string data. We have had specific lecture delivered on clarifying different type of data, string and non-string data.

String data where you cannot do mathematical operations, you can at best get the frequency distribution. So, this is how we are going to make the dictionary file. Now, I am going to do it on the screen in front of you as well.

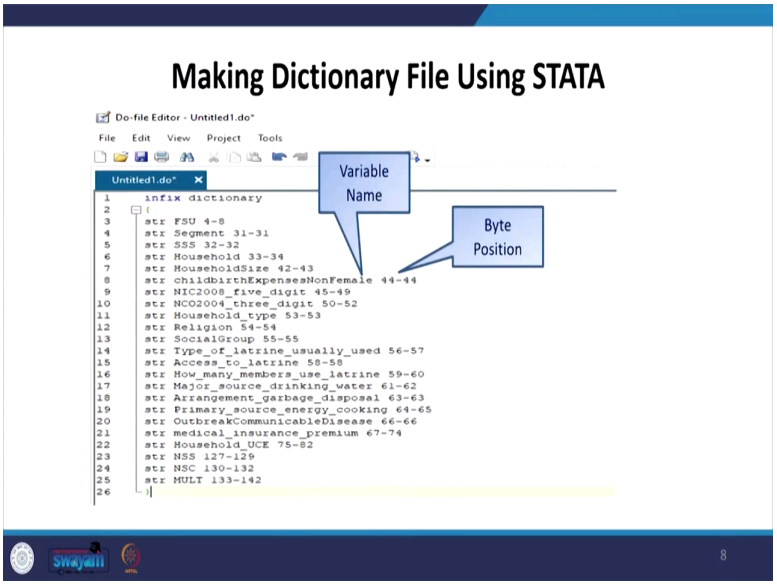
(Refer Slide Time: 26:39)



(Refer Slide Time: 26:41)

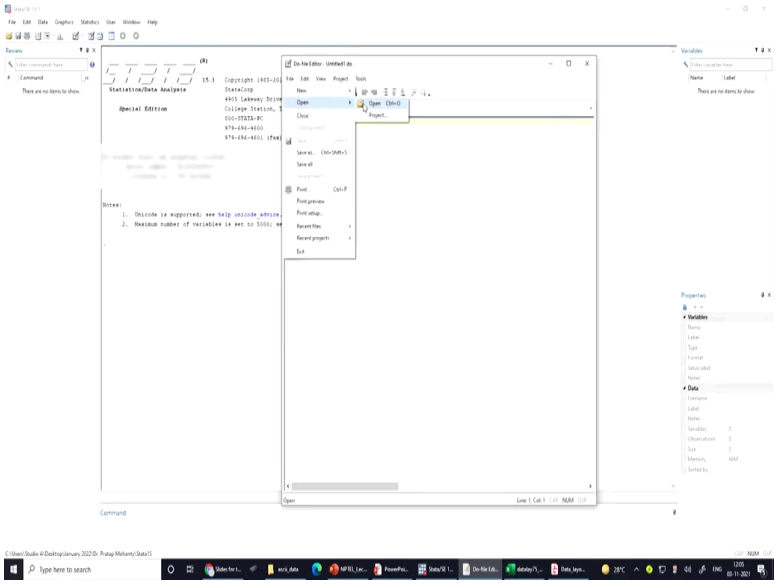


(Refer Slide Time: 26:43)



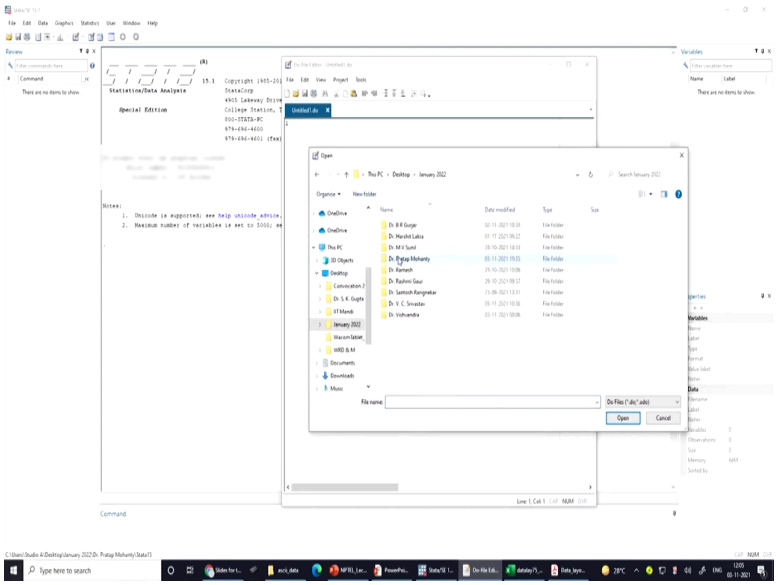
Let us open in STATA one dct file, one do file.

(Refer Slide Time: 26:52)

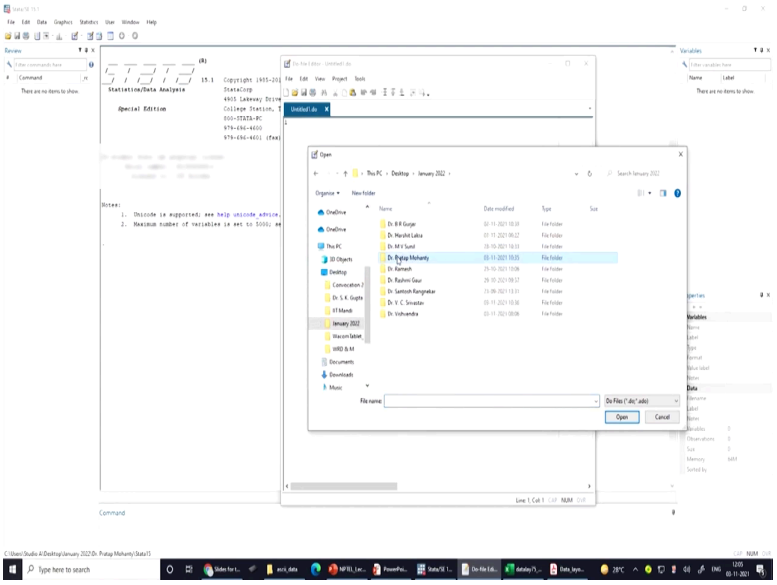


We will open our one sample do file, then I will guide how you can do it on your own. This is the page you must open first; then you open a file which we have already stored it, but you can easily do it.

(Refer Slide Time: 27:07)

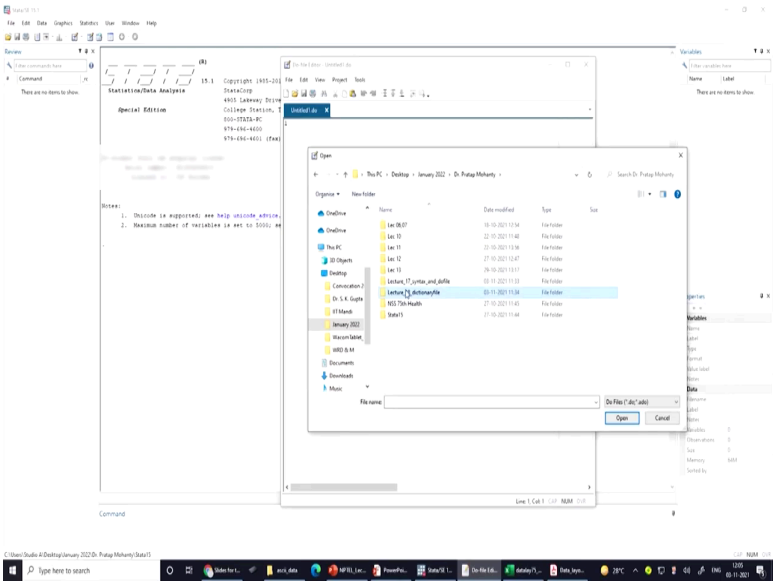


(Refer Slide Time: 27:10)

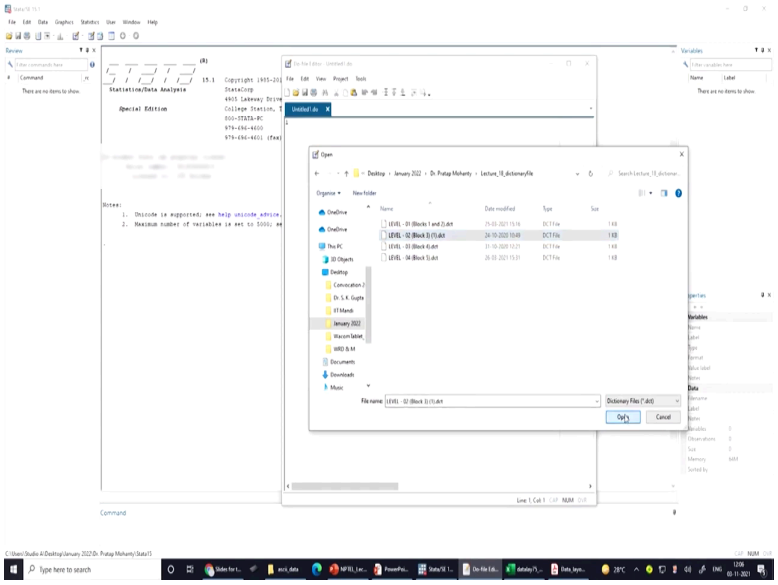


We have deliberately stored it, so that it will be less time consuming.

(Refer Slide Time: 27:11)

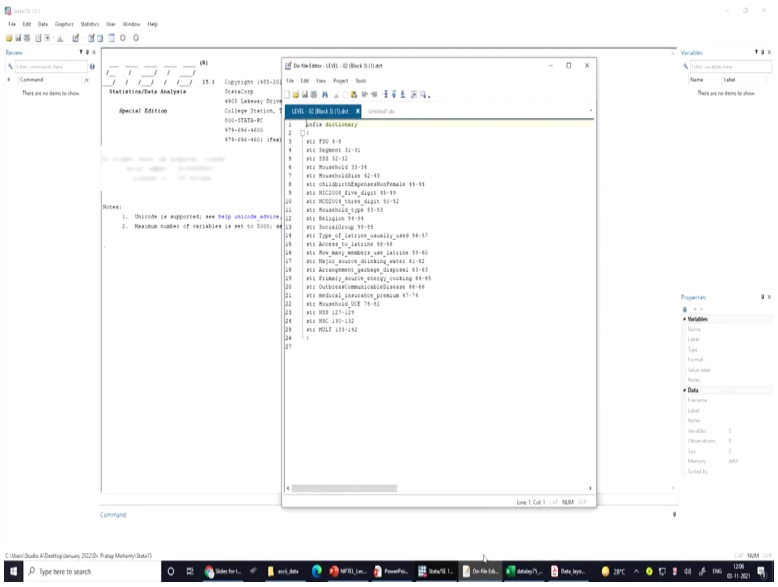


(Refer Slide Time: 27:14)



And mistakes will be lesser and systematic guidance could be given. From here you have to open dot dct format.

(Refer Slide Time: 27:28)



Let it be any file for example, block 2. So, level 2 file is actually explained over here. Like where to start it? I will specify that you start with infix type, exactly as infix, then dictionary; the correct spelling must have been there, then on the next line this bracket has to be also specified.

And now onwards you are supposed to start with the variable and their byte positions. At the end the bracket must be closed. So, starting with what you type it on your on the screen str. Why we are seeing str? Str indicates your string variable.

Why we are making this as a string? Because of the fact that we may require some of the variables to combine and deal and to convert that to be for us as a primary key. So, if the variables are separated, different variables are required to be combined to develop a common id. But if these variables are extracted with numeric format, if we simply combine those variables that, those will be added with their numeric value.

But if those are in string format; like for example, your Aadhar card number or pan card number; it has numerical number, it has coding, it has some asterisk mark, there are so many things entered. But if suppose some codes are given like 2 or 3; 2 does not mean it is by value it is double or 2, 3 does not mean 3 is higher than that of 2.

3 might be indicating a feature, a character; like 2 for example, in gender if you are saying 2, 2 may indicate female. So, 2 is not higher than that of 1. So, that is why string is required; if you are getting the extracted values in string, later on you can easily convert it with another command called destring.

At this moment let us make the variable as string. The first variable name here, it is FSU. Why FSU? Why it is 4 FSU, then segment SSS, household? We know that from our data layout, you just have a check once; what is your common primary id. Why common primary id is required? I told you that this is required for extraction and for merging of different blocks; we have 13 different levels mentioned.

We know that different models are in fact useful for our analysis. While merging all those information, we have to map it which entry is similar in other blocks or in other levels. So, we have to confine the particular identification in every cases. So, what is those identification?

Those identifications are explained with FSU number, segment number, then SSS number, second stage stratum number that is provided by NSS, then household number. Since we are extracting households, so household number is required. If you are extracting the individual one, then personal id is also required, at this moment we are in fact explaining the household

information. So, the do file we specified as per the data. So, our primary key id will be FSU, its byte position is given.

Here byte position is guided i.e., FSU is from 4 till 8; 4, 5, 6, 7, 8, there are 5 positions occupied by FSU. So, this is 4 till 8 is defined. Or even in the first block it has guided about FSU very correctly, FSU is 4 till 8, so that you can as easily find it.

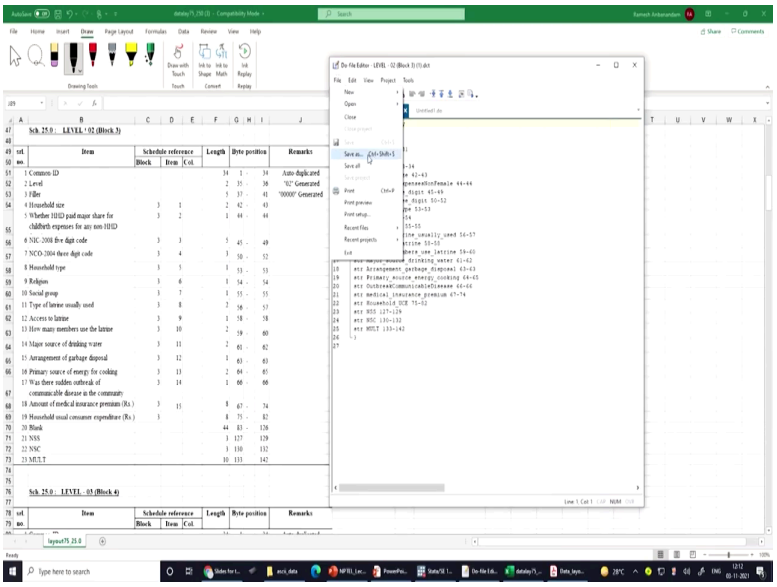
Then next after doing that all those basic ids, FSU segment number 31 till 31; I think it is already given 31 till 31, only one position occupied. So, we have specified correctly. So, household number is 33 to 34; I think household number 33 it is 2 position occupied, so 33 and 34 is occupied.

So, what exactly is important for you is that, you should actually very clearly specify their byte position. Once you define the correct byte position of its primary key; like primary key id I have already guided to you, till household 33, 34 both position of we have defined very clearly. Once again, I am opening the layout.

Rest of the entries are in fact, the variables, the information about the household; like household size, household type, their religion, social groups, access to laterin, major source of drinking water etc. At the end there are multiplier characters entries, these are helpful for developing a weight of the variable.

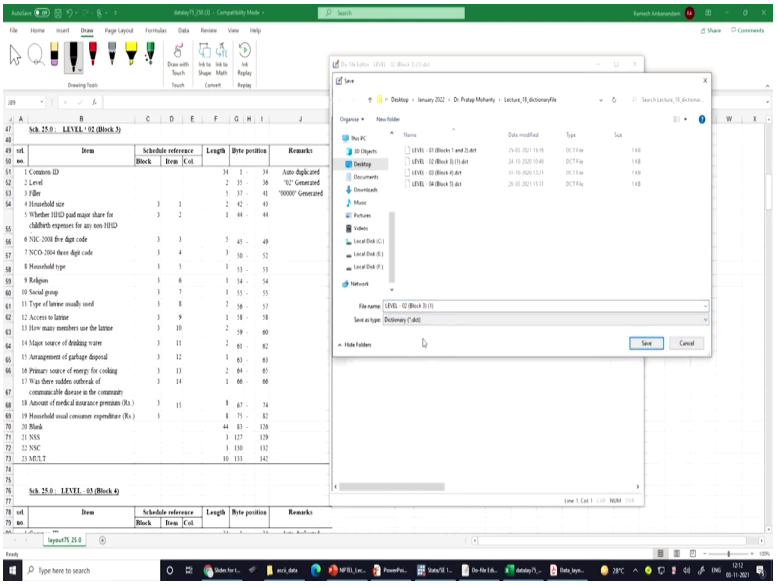
So, these are entered with their correct appropriate byte space. Make sure that you have entered a bracket and a bracket that should be closed; though since we started with a bracket, the same bracket has to be closed at the end.

(Refer Slide Time: 33:57)



Now after doing so, you have to save it; file save as with a name, then with the dct file.

(Refer Slide Time: 34:00)



We have to open it with a dot dct file and save it; that is all you have to do it and your dct file is ready for work.

And how we are going to use it, in our other operations I will guide you throughout; we have also highlighted in green colour about creating dct file for your use. I think I have already guided you, I need not stress much; these are sufficient for your understanding. If still there

are something missing, I am sure I will clarify in my other lectures those are going to be followed. So, that is all for today, looking forward to your participation in the next class.

Thank you.