IIT Madras

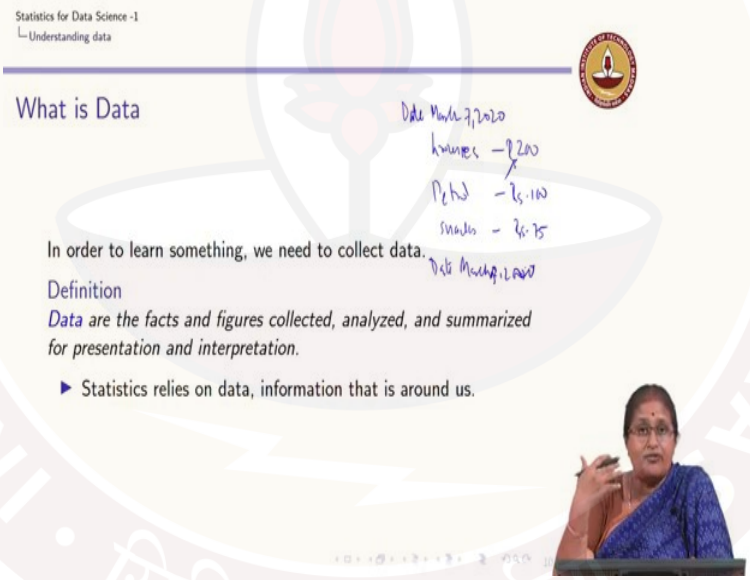
ONLINE DEGREE

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture - 04
Introduction and Types of data Part – 2

So, the next part we are going to understand data. This is extremely important for us because statistics relies on data and when we say data it is information that is all around us, whether we are formally doing a statistical analysis or not; all of us are either creating data or contributing towards collection of data or we are collecting data ourselves. There are so many times when starting with the simple household accounts data which we keep every day.

(Refer Slide Time: 00:52)



We decide on how much we spend on say every day, we have people who maintain accounts and accounts could be of the kind that I come every day, I write down a date and then afterwards I say groceries 200 rupees, then petrol rupees 100.

I put the date say I put a date which is say March-7-2020 and then afterwards I will say I might have spent something on snacks; I have spent rupees 75 on snacks. Again I go back and I come back after 2 days, I do March 9-2-2020. I again calculate my data, I have again data; all this is also a data. We all collect data all the time. There is a lot of

data and we are contributing to data also every time we click the button or a keyboard or the mouse, we are generating some data.

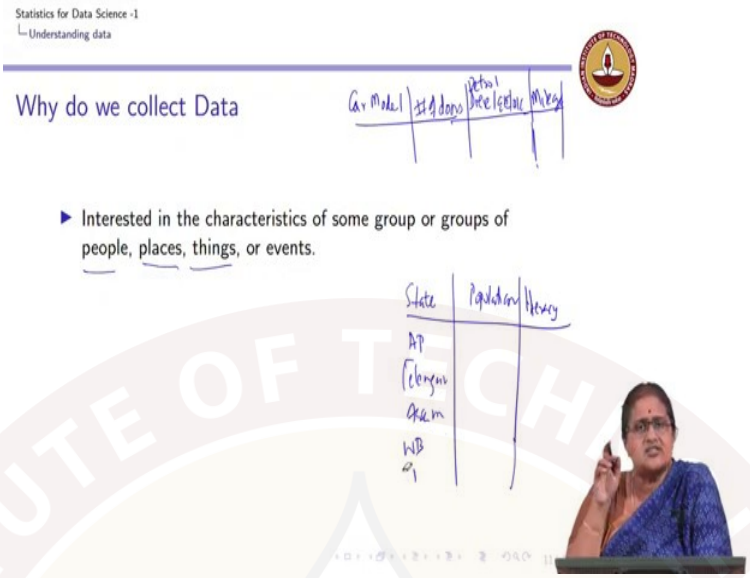
So, as I said the definition of statistics has changed drastically over the years. So, has the nature of data. What data would have meant about 50 years back is just about numbers and categorical data. Today people talk about social media analytics, multimedia analytics, text analytics and there is so much data; even a comment on a YouTube video or a multimedia video or a photo is data. You can see that comments that come on a product e-commerce portal that is data.

So, there is data which is being collected as we speak. There is so much data that is generated there is so much data that is being collected and there is so much data waiting to be processed into meaningful information. That is the purpose so and whenever we want to do a statistical analysis, we rely heavily on data.

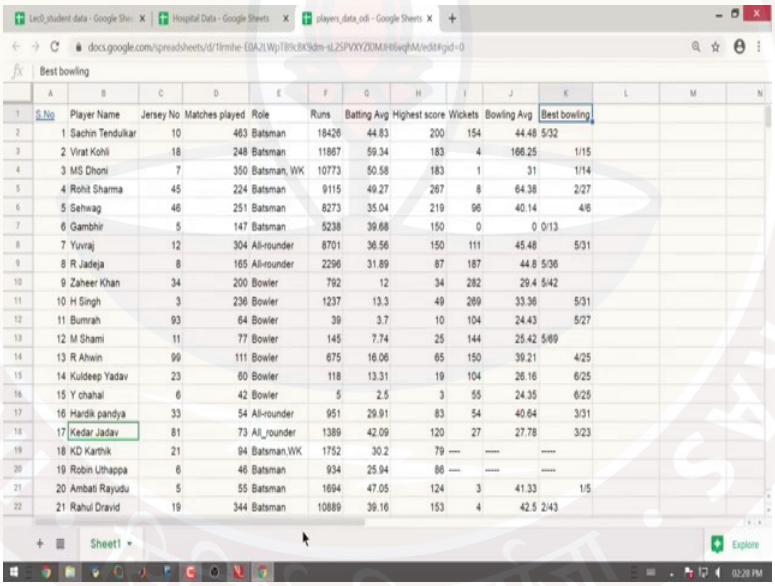
So, first let us define what is data, very simply put data is just facts and figures collected; by facts, it could be numerical, it could be any type. So, I just said that the comments on a multimedia or a video or anything on the internet or a product all of this contribute to what we call data.

So, what is data? Data is fact; it is what is there. We want to summarize this data we want to analyze this data for presentation and interpretation. It is very wrong to say that I want my data to tell something. You do not want the data to say anything, you in fact, the data is a fact you use data to extract information what the data is telling for interpretation purposes. So, this is the knowledge we need to understand what this data is all about and in this module, we are going to understand what is data.

(Refer Slide Time: 03:59)



(Refer Slide Time: 04:08)



Why do we collect data? Now we go back to our cricketing data set why do we collect why is this data collected at all? What are the, what is the purpose of this data? We might want to answer questions that given this data, we would like to answer questions that who is the person who has played the highest number of matches. This is a very valid question to ask or who has the highest batting average or we would want to know who is a person who has taken the highest number of wickets.

So, you see that or I would want to know here of course, I have only Sachin Tendulkar has played matches 463 matches, but then afterwards I would want to seek data about every match Tendulkar has played of this 463 matches to see that how has his batting performance been over every match and where he has played that match to see whether there has been a difference in his in house or in country batting performance to out of country batting performance. I might want to do that with every player. So, as I speak, you can see that we are generating a lot of questions and for to answer all these questions what we need is data.

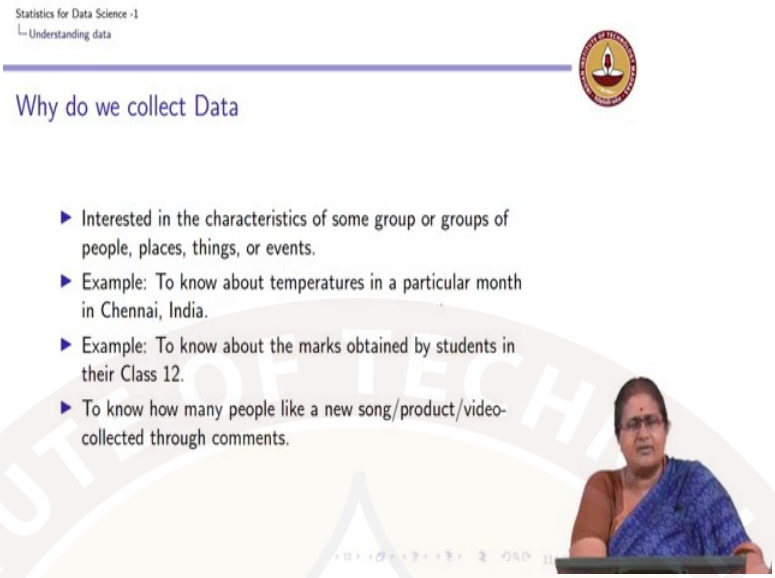
So, what is the why do we collect data? We collect data the primary reason why we collect data is we are interested in knowing about the characteristics of groups; it could be groups of people, it could be places, it could be things, it could be events. Notice that we are not always interested only in people.

For example, I could have just a collection of data wherein I have a car model, number of doors of that car, whether it is diesel, whether it is petrol, whether it is an electric car. So, here you see I have absolutely no people involved or group of people involved, how many then what is the mileage of the car, whether it is a sedan or whether it is a hatchback; all these things are something which I am going to collect.

So, you see immediately whenever I want to talk I am saying groups of people it could be things, there could be a data which I am collecting wherein I have this state and the population, literacy rate. So, the state Andhra Pradesh, Telangana, Assam, West Bengal, Tamil Nadu; I could note down the population and what is the literacy rate.

Here you see I am interested in knowing about the states so, it could be places; it could be anything. So, when we are restricting ourselves it need not be only to people. So, why do we collect data? We collect data whenever we are interested in some characteristic or attribute and we seek data to answer about these characteristics or attributes.

(Refer Slide Time: 07:30)

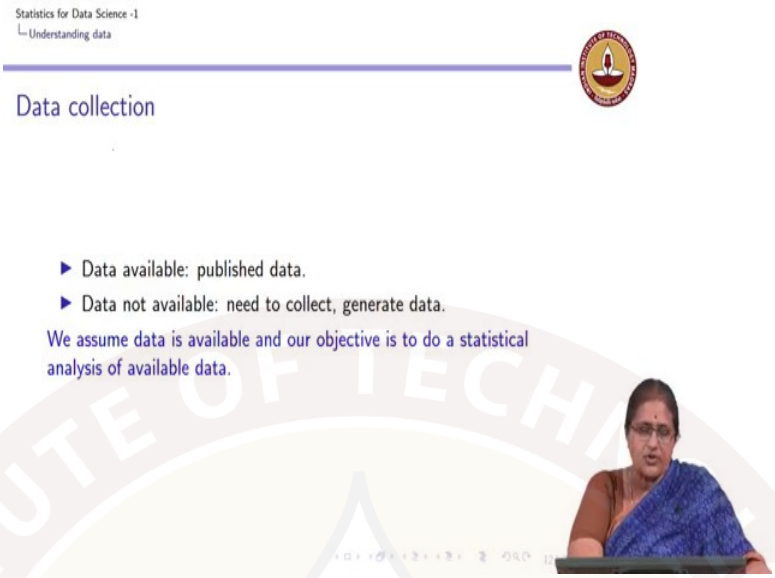


Example, I would also know want to know about the temperature in a particular month in Chennai. So, Chennai is a place again I am not interested about people here, I would want to know about the marks obtained by students, I might also want to know how many people like a new song. These days any internet or anything that is streaming over the internet you will find likes and dislikes. So, you might want to know how many people like a new song, new product, new video. This is collected in an entirely different way through comments.

So, as we speak we see that the data is when I talk about temperatures, you can see that there is a way I collect this data. When I talk about marks, there is a way I collect the data, marks could be either percentages or it could be grades. When I talk about marks, marks need not always and nowadays most of the boards and colleges do not give percentage marks. There are lot of people who have switched over to grades or letter grades to evaluate students.

So, you can see even that constitutes data and comments; comments is also data, but that is a completely different form of data, it is textual data. So, the minute we talk about data we see that data is all around us. We need to understand that we are collecting data to actually because we are interested in characteristics of groups or people or events. So, this is why we collect data.

(Refer Slide Time: 09:26)



The next thing is, where do we collect this data from? Where do I get this data from? To answer this question, you can see that either you go and collect the data; you need to collect data and generate data or there is already data available, there is published data which is available everywhere ok. One site which you can always go and look at most of the governments publish their own data sites.

(Refer Slide Time: 09:57)

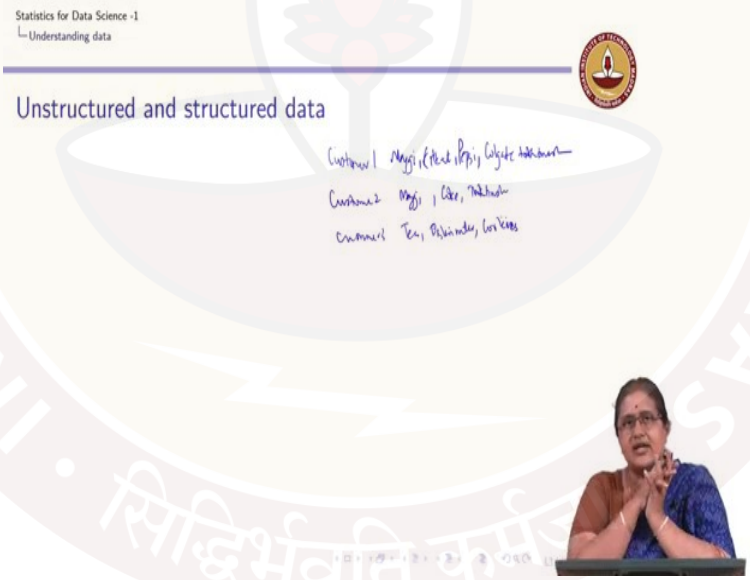
So, if you look at; so, this is a site data dot gov. It is an open government data platform India. So, this data site almost gives about all the data that is collected at a government

level. So, you can see that you have drinking water and sanitation, health data, you have economy data, the transport data, education data; so there is a data available here.

So, the key idea here what I want to convey is data is either you go and collect data or you have published data you can work on any data. If the questions you are seeking to answer needs data that has to be generated or collected; you have to go and collect the data.

This course is not going to lead you to understand how to collect this data or generate this data, but we assume that the data is available to us and our objective is to do is statistical analysis of available data. That is the purpose or the objective in this course. Nevertheless if you are seeking answers to questions for which data is not available, you need to understand how to collect this data in a structured or a scientific way how to generate this data. This will also be taught to you in due course.

(Refer Slide Time: 11:41)

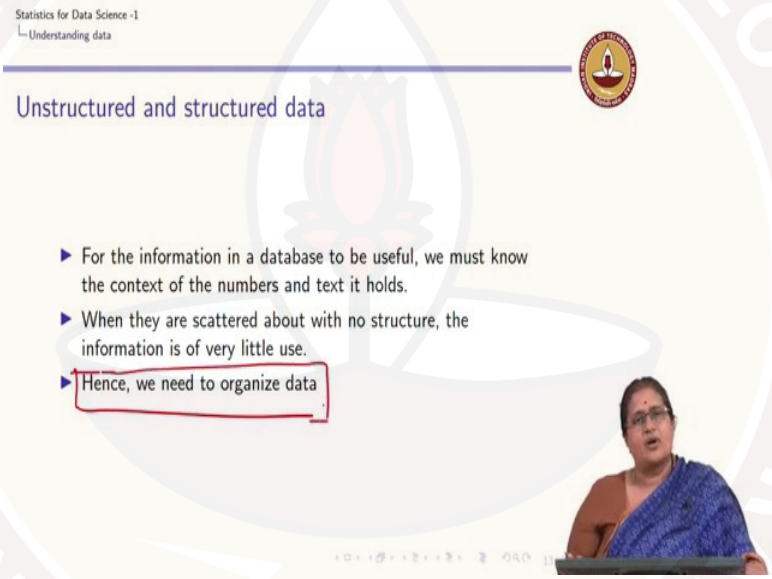


Now, when I come to data as I said that suppose I have a file; I have a file which is of this kind I just ask a person in a perhaps a retail market and I ask him what are the things that have been sold. And he comes up with a data of this kind customer 1 bought Maggie, KitKat, Pepsi and perhaps some Colgate toothbrush. Customer 2 bought Maggie, Coke, toothbrush. Customer 3 has bought say tea, then Bisleri water, they bought some cookies.

So, this is the data the person sitting at a retail counter is just collecting and if they present to you something of this kind, you cannot make any meaning out of this data which is presented to us. In a sense that this is data nevertheless this is data to you, but can we make any meaning out of this data? Suppose imagine this person who has collected this customer data is giving us similar data for 50 customers. Ok?

So, immediately this is data, but it is in an unstructured form. We have not given any or it is in an unorganized form this is data nevertheless, but it is not in a very structured or in an organized form. So, in this course we are looking to only analyze data that come to us in a structured form.

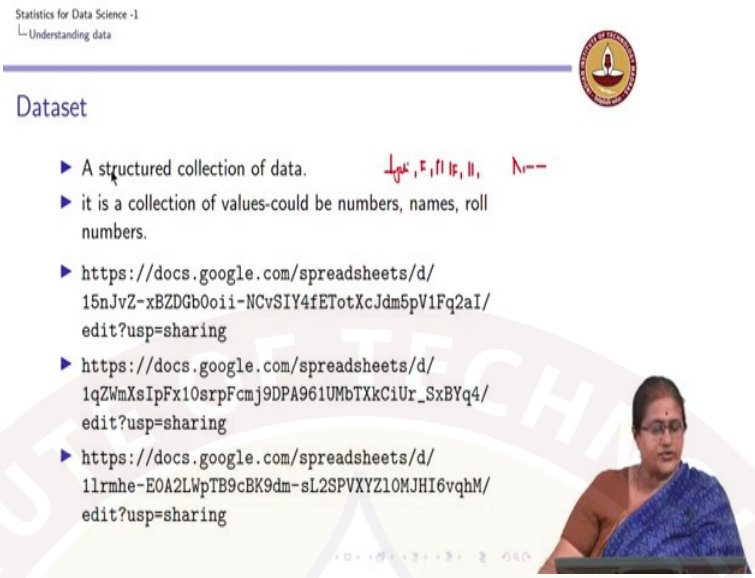
(Refer Slide Time: 13:48)



So, for information of a data to be useful we must know the context of the numbers and text it holds. When they are scattered just as the example I described earlier, it is with no structure the information is of very little use, but; however, I need to organize data. So, the compelling need for me now is to organize data. This is the most important thing which I need to do.

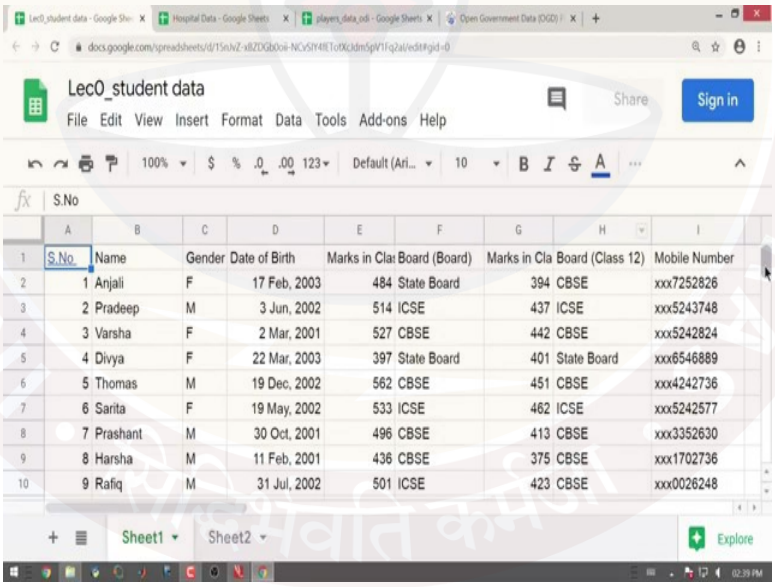
What do I mean by organizing data?

(Refer Slide Time: 14:25)



So, I look at a structured collection of data.

(Refer Slide Time: 14:38)



S.No.	Name	Gender	Date of Birth	Marks in Cla	Board (Board)	Marks in Cla	Board (Class 12)	Mobile Number
1	Anjali	F	17 Feb, 2003	484	State Board	394	CBSE	xxx7252826
2	Pradeep	M	3 Jun, 2002	514	ICSE	437	ICSE	xxx5243748
3	Varsha	F	2 Mar, 2001	527	CBSE	442	CBSE	xxx5242824
4	Divya	F	22 Mar, 2003	397	State Board	401	State Board	xxx6546889
5	Thomas	M	19 Dec, 2002	562	CBSE	451	CBSE	xxx4242736
6	Sarita	F	19 May, 2002	533	ICSE	462	ICSE	xxx5242577
7	Prashant	M	30 Oct, 2001	496	CBSE	413	CBSE	xxx3352630
8	Harsha	M	11 Feb, 2001	436	CBSE	375	CBSE	xxx1702736
9	Rafiq	M	31 Jul, 2002	501	ICSE	423	CBSE	xxx0026248

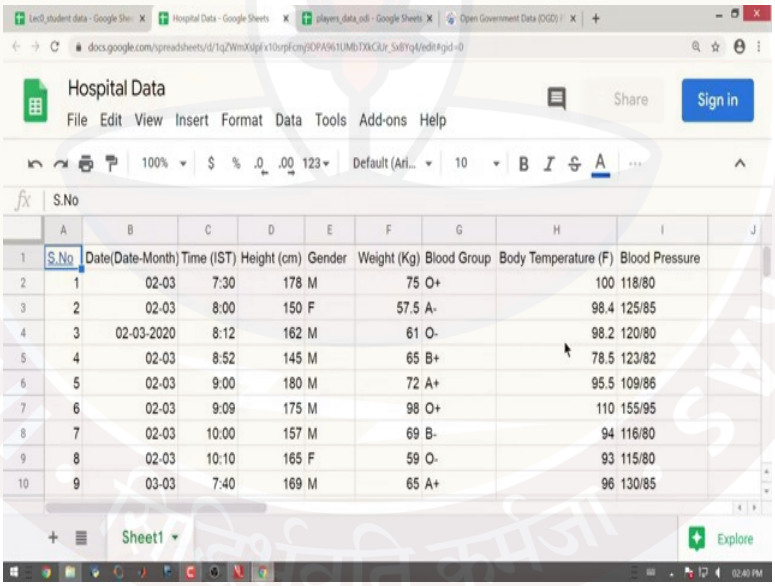
So, if you look at a structured collection of data, we could see this is the data set which you are already in this is the data set we were already looking at when we will this is the data set we have already looked at in the last when we introduced the course. So, here you can see again this is a hypothetical data set. Imagine if this data set were collected by a person when every student was entering the college and all that the person was doing was as a person was entering the college wrote down Anjali, female, board, what is the

board ICSE, marks obtained 98 and all of that then afterwards Ramu, male, and all of this if that person had done this kind of a data collection, then again it would have been unstructured.

So, what we look at is we are trying to give a structure to this data in terms of what we refer to as a data table and in this course, we are going to largely use Google sheets to analyze our data. Hence we want to put this data or there are many ways of tabulating your data.

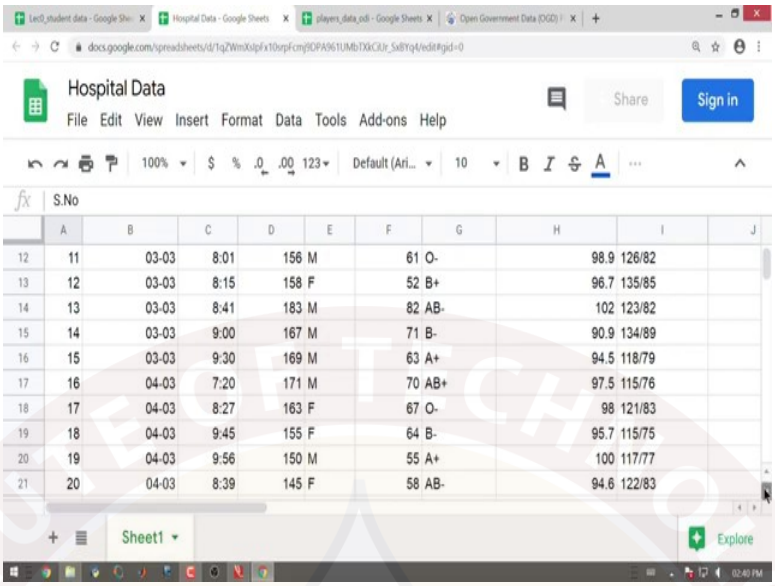
We are going to now describe what is the way we are going to use to analyze data in this course and mainly we are going to use spreadsheets this Google sheets to analyze a data. So, we are going to understand now how we are going to structure or organize data in a spreadsheet, how would we go about it that is the next thing which we are going to organize.

(Refer Slide Time: 16:24)



S.No	Date(Date-Month)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	02-03	7:30	178	M	75	O+	100	118/80
2	02-03	8:00	150	F	57.5	A-	98.4	125/85
3	02-03-2020	8:12	162	M	61	O-	98.2	120/80
4	02-03	8:52	145	M	65	B+	78.5	123/82
5	02-03	9:00	180	M	72	A+	95.5	109/86
6	02-03	9:09	175	M	98	O+	110	155/95
7	02-03	10:00	157	M	69	B-	94	116/80
8	02-03	10:10	165	F	59	O-	93	115/80
9	03-03	7:40	169	M	65	A+	96	130/85

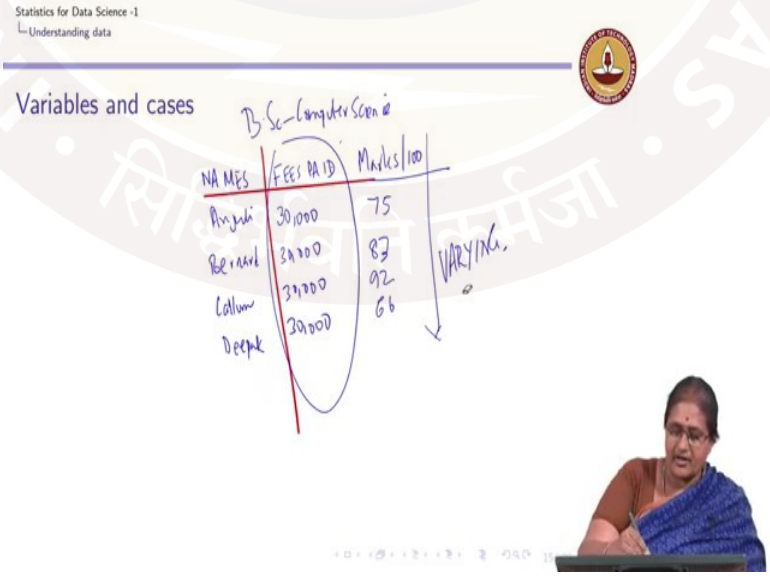
(Refer Slide Time: 16:32)



	A	B	C	D	E	F	G	H	I	J
12	11	03-03	8:01	156	M	61	O-	98.9	126/82	
13	12	03-03	8:15	158	F	52	B+	96.7	135/85	
14	13	03-03	8:41	183	M	82	AB-	102	123/82	
15	14	03-03	9:00	167	M	71	B-	90.9	134/89	
16	15	03-03	9:30	169	M	63	A+	94.5	118/79	
17	16	04-03	7:20	171	M	70	AB+	97.5	115/76	
18	17	04-03	8:27	163	F	67	O-	98	121/83	
19	18	04-03	9:45	155	F	64	B-	95.7	115/75	
20	19	04-03	9:56	150	M	55	A+	100	117/77	
21	20	04-03	8:39	145	F	58	AB-	94.6	122/83	

If you look at another data set here you can see that this data set has collected over some 9 patients; sorry it has collected about 20 patients who are entering a diagnostic center over a period of time 7:30, 8:00, 8:12; the height, gender, blood group, body temperature and blood pressure. So, again you see that this gives us a sort of tabulated data. Now we are going to understand what we mean by a structured collection of data and that is what we are going to focus on now.

(Refer Slide Time: 17:11)



When I mean by a structured collection of data to form a data set, I first need to understand, what is a variable.

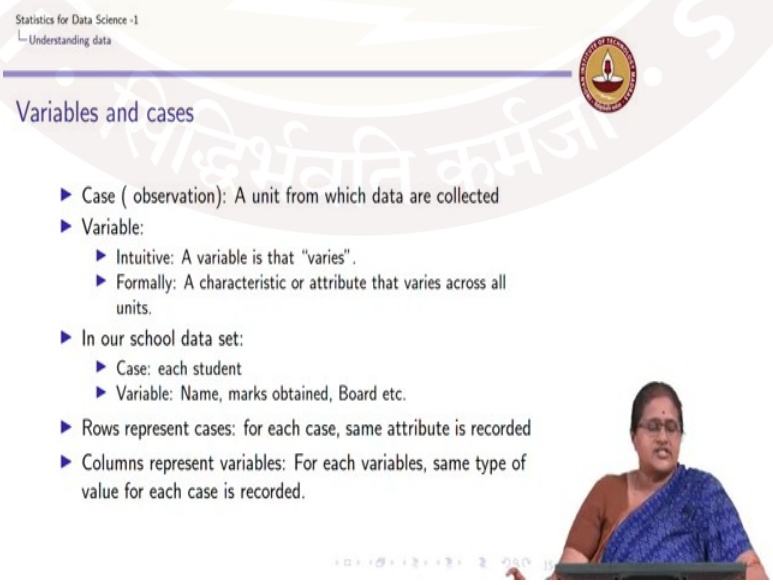
Now, suppose again I have a set of data which looks like this I have name and I am looking at the fees paid, the course is just a BSc course say BSc Computer Science course, assuming. This is again hypothetical data. I have names again Anjali, Bernard, Callum, Deepak etcetera.

When you look at the fees paid again these are hypothetical numbers. If I am looking at 30000 all in INR 30000, 30000, 30000; if I am looking at fees paid and the next thing I am going to look at is what are the marks obtained out of 100 I could have 75, 83, 92, 66.

Now, when I look at this table, you can see that when I look at fees paid all of them are paying the same fees. So, there is absolutely nothing that I want to ask when it comes to fees paid. Everybody is paying the same fees, nothing is changing it is a constant along all these people. But whereas, I look at the marks I can see Anjali has obtained 75, Bernard has obtained 83, Callum has obtained 92 and Deepak has obtained 66; in other words this is varying.

I have a concept of variability there. So, when we look at what is a variable, the answer is very simple here again that we are just introducing at a basic level what is a variable. So, you can see that I can define a variable as the following.

(Refer Slide Time: 19:29)



I can define a variable as something that "varies" and formally it is a characteristic or attribute that varies across all units.

Now, let us understand what is a unit and a what is a variable in each of the data sets I have described so far. Now if we go to this data set, you can see that name is a variable in a sense, gender is definitely a variable it is not taken from a if I were taking this data from a men's-only college or a girls-only school, then this would not have been a variable; it would have been a I would have had everybody from the same gender. But here you can see that it is variable, the marks obtained is variable.

So, is the date of birth not everybody was born on the same day it could be likely again. It could be very likely if the year of birth was taken perhaps it was not varying a lot, but again it would definitely there would be in some light amount of variability there; the board, the mobile number will come and see.

So, here if you look at it the way we have defined name, gender, date of birth, marks, the board and all of them are variables whereas, Anjali, Pradeep, Varsha, Divya they are all cases or observation. On each case on each case, I have each variable recorded for each case for Anjali I have each variable recorded.

Similarly I come to the hospital data. The variable is time the date; if you look at the date you see, it was over the same day. So, here we can see that date is not varying, but here yes it is third. The first eight observations if my data is a subset and I am looking only at the first eight observations, it was taken on the same date. So, that is not varying whereas, time of entries varying. This is also data; height varies, gender varies, blood group varies, so is the blood pressure and body temperature.

So, you can see that in this case the person I have not noted down their number, but I can call them the person 1 is recorded at time of arrival, height, gender, weight, blood group, body temperature and blood pressure. Similarly each person who enters the system is recorded on each of the variables.

Similarly in the players data set, I have jersey number. Now interestingly you see jersey number also no two players have the same jersey number and you see that everybody has a different jersey number, the matches played, the role, country, here that could also be a

variable, but since this is only India then country is not a variable; highest score, wickets, bowling average.

Now, interestingly in this data set, you find some data of this kind. You can find in this there is a data which is telling 0 and there is also a data which is just showing dashes. Now what does this mean? You see that Gautam Gambhir did not take any data, but he has bowled, ok? So, the value is 0 whereas, when you look at the data of Dinesh Karthik or Robin Uthappa, you see that they do not have the data that is available for giving them any bowling statistics at all.

So, in this case even though I am collecting data, there it could be quite possible that the data which I am seeking a subset of variables which I am seeking might not be available for every unit as we have seen in this case, ok?

So, this is what we refer to this data is not available, nevertheless these people are a very much part of the data set. They do have a batting average ok, but their bowling averages are not available. What I want to emphasize at this point is this non-availability of data is different from a data taking a value 0; it means a lot. This 0 is taken even though a person has bowled 13 overs whereas, here the data is not available because these people have not bowled. We cannot take it as 0, it would mean a completely different story at this point of time.

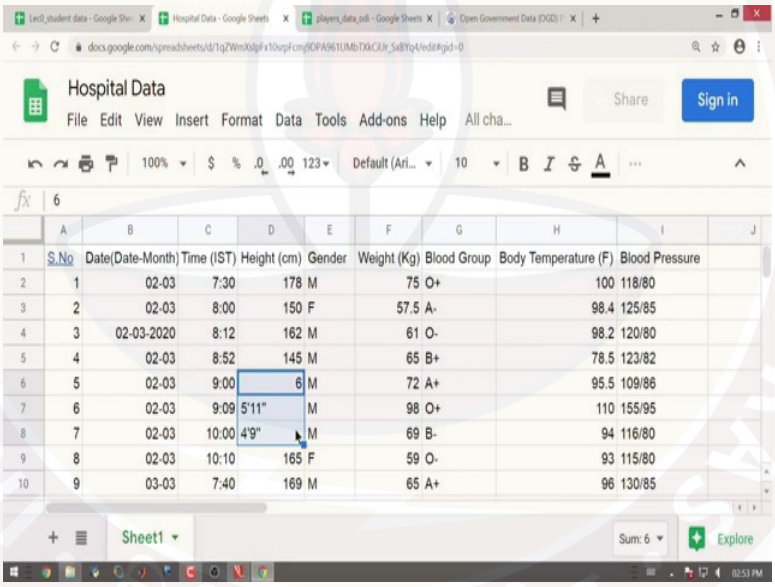
So, when we look at data the first thing which we need to understand is what is an observation and what is a variable. So, intuitively a variable is that varies formally it is a characteristic that varies across all units. If the characteristic is available for that unit, as we saw in the cricket data set that characteristic is not available for certain players, the characteristic of bowling averages.

In our school data set, each student was a case the variable was name, marks obtained, board etcetera, rows represent cases for each case same attribute is recorded either it is recorded or we say not available. In the cricketing data set, the attribute of bowling, bowling average was not available for certain players. So, even though we record as not available if that attribute is not available, we record it as 0 if it is a value 0, there is a difference between a value 0 and not available.

Columns represent variables and for each variable the same type of value is recorded. What do I mean by same type of value? Again let us go back to our hospital data set. In this hospital data set, there are two variables here, one is which is the height variable and one is which is the weight variable.

Next to height, you see a centimeter which is written and next to weight you can see that a kilogram is written. Now suppose and again you look at it at the data collection starts at 7:30 and ends at 10 O'clock on 2nd March. If I have people who are working in shifts of 2 hours duration, a person who starts at 7 ends at 9 and a person who starts at 9 goes up to 11. The person who comes in for a shift at 9 decides to take the person's height and mention it in feet. So, he would she would or he or she would be mentioning 180 centimeters as 6 feet rather than 180.

(Refer Slide Time: 27:25)



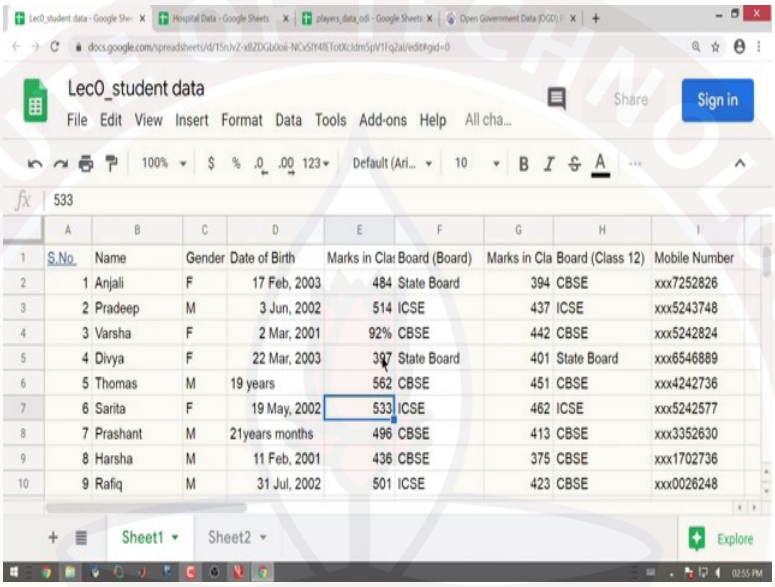
S.No	Date(Date-Month)	Time (IST)	Height (cm)	Gender	Weight (Kg)	Blood Group	Body Temperature (F)	Blood Pressure
1	02-03	7:30	178	M	75	O+	100	118/80
2	02-03	8:00	150	F	57.5	A-	98.4	125/85
3	02-03-2020	8:12	162	M	61	O-	98.2	120/80
4	02-03	8:52	145	M	65	B+	78.5	123/82
5	02-03	9:00	6	M	72	A+	95.5	109/86
6	02-03	9:09	5'11"	M	98	O+	110	155/95
7	02-03	10:00	4'9"	M	69	B-	94	116/80
8	02-03	10:10	165	F	59	O-	93	115/80
9	03-03	7:40	169	M	65	A+	96	130/85

So suddenly your data set would start appearing after a certain point as 6. You would start looking at a data set. So, this would be 6; this might be some 5 feet 11 inches and things like that 4 feet 9 inches.

Now, you see immediately and then at 10 O'clock another person comes and the again restart. So, you see that these three data units even though they are measuring height, there is no consistency in the units that have been used. Now this as we look at a data set a primary glance of the data set itself tells us there is some problem with the data set.

Whenever we measure data, measuring a variable if it has units; we had to be consistent about the units we are using across all the observations. And that is what we mean by saying that columns represent variables and for each variable, the same type of value for each case is recorded. Again by type of value what do I mean? I cannot go back to this data set and where for example, in my data set here I have date of birth which is 17 February 2003.

(Refer Slide Time: 29:01)

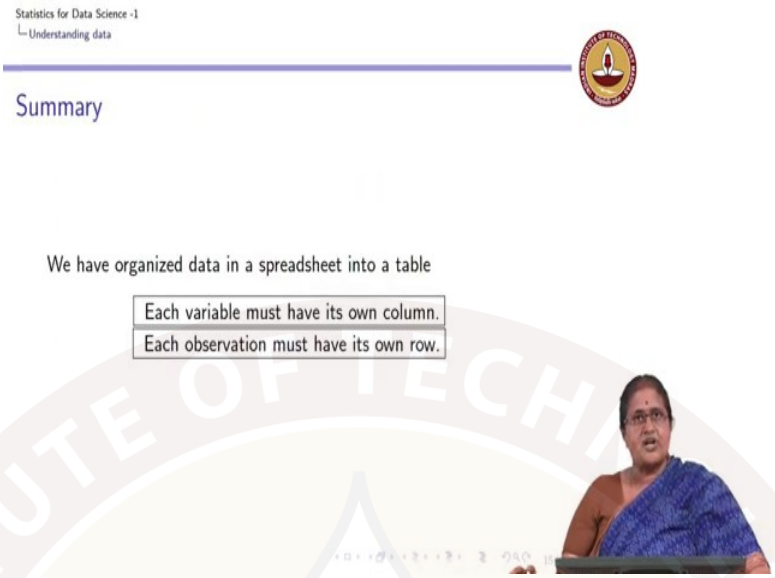


	S.No	Name	Gender	Date of Birth	Marks in Cla	Board (Board)	Marks in Cla	Board (Class 12)	Mobile Number
2	1	Anjali	F	17 Feb, 2003	484	State Board	394	CBSE	xxx7252826
3	2	Pradeep	M	3 Jun, 2002	514	ICSE	437	ICSE	xxx5243748
4	3	Varsha	F	2 Mar, 2001	92%	CBSE	442	CBSE	xxx5242824
5	4	Divya	F	22 Mar, 2003	397	State Board	401	State Board	xxx6546889
6	5	Thomas	M	19 years	562	CBSE	451	CBSE	xxx4242736
7	6	Sarita	F	19 May, 2002	533	ICSE	462	ICSE	xxx5242577
8	7	Prashant	M	21years months	496	CBSE	413	CBSE	xxx3352630
9	8	Harsha	M	11 Feb, 2001	436	CBSE	375	CBSE	xxx1702736
10	9	Rafiq	M	31 Jul, 2002	501	ICSE	423	CBSE	xxx0026248

Now, suddenly I might not want to change for some people; I might not want to come and say 19 years or 21 years 3 months. Technically speaking both of them are capturing in some sense if I know the date of birth, I can compute what is a age; if I know the age, I can compute what is the date of birth. But they are not the correct way of representing data.

So, when I am computing or collecting data especially in the format I wanted to I need to ensure the following that the rows represent each case and columns represent variables and for each variables, I ensure that same type of value. The another example let us go back again here, I have marks obtained in the class so, 484. I cannot suddenly say this is 92 percent or something even though that is technically it is an evaluation right.

(Refer Slide Time: 30:28)



So, that is something which we need to take care of that every; so, when I talk about a data set what I need to be very careful and what I need to understand is I have a data set for which I know each variable has its own column, I have defined what is a variable.

If the variable has units, then every observation has its own row and every observation has the variable or each variable is measured for every observation, the units are consistent. I cannot have an observation taking the unit of height which is a variable in centimeters and another observation which is giving the unit of height in feet. I need to have the variable height taking the same type of value for each observation.

So, at this stage we understand what is data and the data set for this course is the data set which would be organized in a spreadsheet. We have shown a couple of examples here. This is the school data organized in a spreadsheet, this is the hospital data which is again organized in a spreadsheet, these are the players data which is organized in the spreadsheet. What we need to understand and remember is the columns represent the variables, the rows represent observations or cases.

For every observation I am marking what is the variable value. If the variable value is not available, I put a not available symbol; I capture the non availability of that particular variable for that observation. So, at this point of time, we have a data set available for us for analysis. So, the next step we need to understand is what do we understand about this variable, how do I classify these observations; that is the next thing.

(Refer Slide Time: 32:34)