INDIAN INSTITUTE OF TECHNOLOGY MADRAS

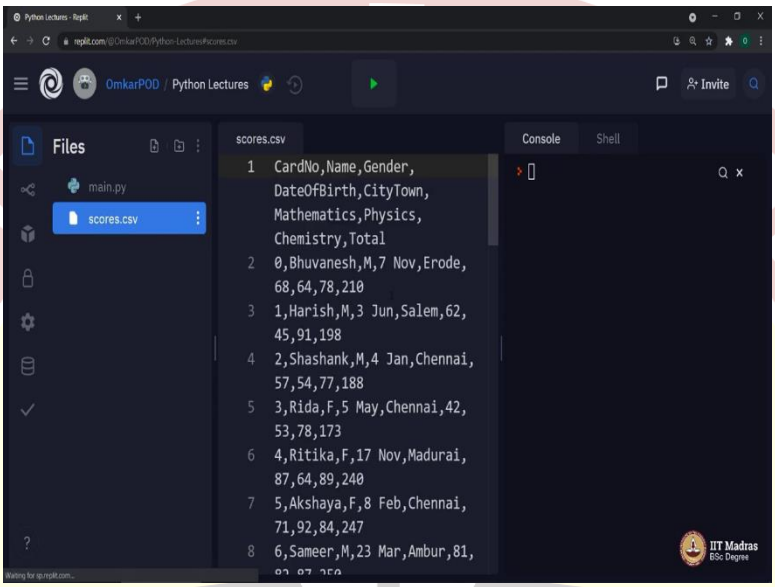
सिद्धिर्भवति कर्मजा

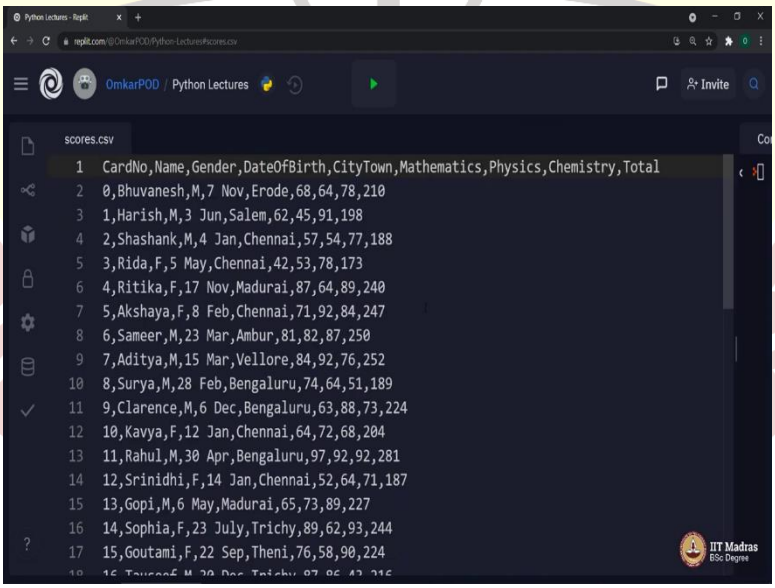
IIT Madras

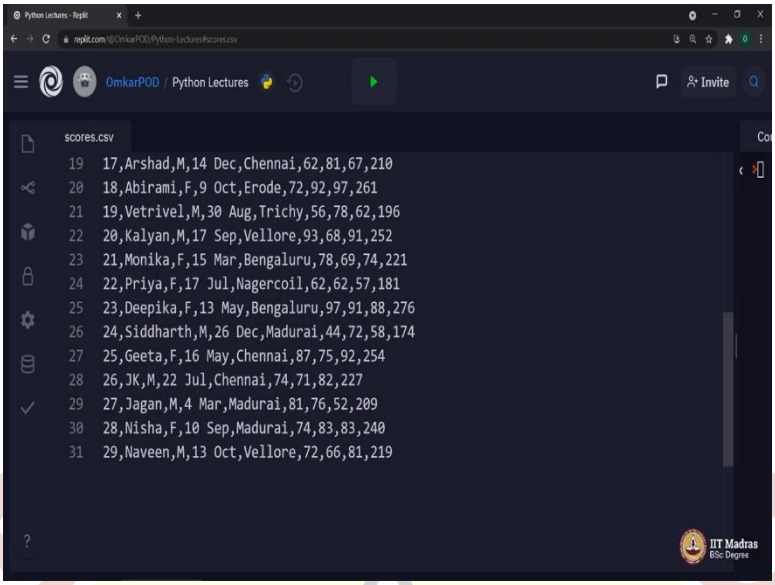
ONLINE DEGREE

Programming in Python
Professor. Sudarshan Iyengar
Department of Computer Science and Engineering
Indian Institute of Technology, Ropar
Mr. Omkar Joshi
Course Instructor
Indian Institute of Technology, Madras
Why Pandas

(Refer Slide Time: 0:16)







Hello, python students. In this lecture, we will discuss about Pandas. Pandas is an external library, which provides high performance data manipulation and analysis tool. We all are aware about the popularity of python in the field of data analytics and data science. And one of the most important reason is Pandas. It makes the python programming language even more powerful.

Therefore, in this lecture, I will try to demonstrate the power of Pandas using some examples. In order to do so, let us consider one input file. I have added one file called scores dot CSV. This file holds the exact data set, which we all know from computational thinking course. If you remember, initially, we started with cards. Later on, we converted those cards into a table, and every column in the table had a specific heading.

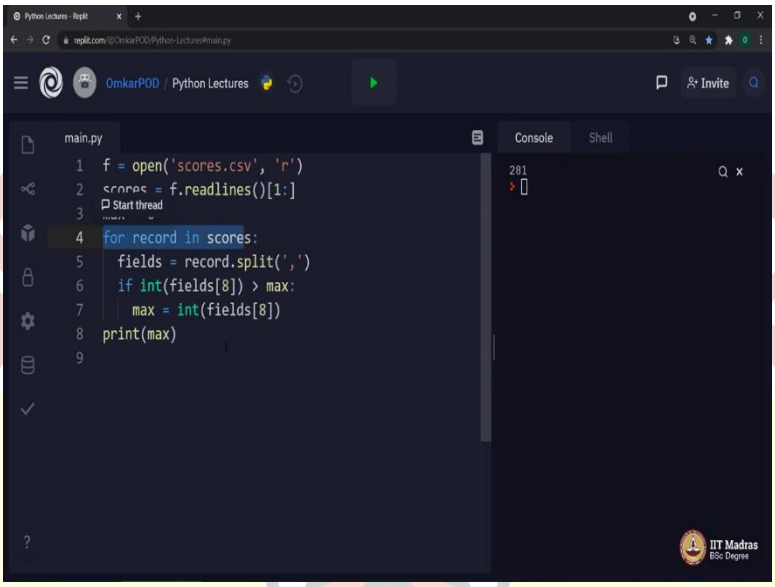
Those headings are given over here in the first row of this particular file, card number, name, gender, date of birth and so on till total. Whereas, from line number two onwards we have our actual data from scores data set. And if you remember, we had 30 such cards, that is where the data is from 0 to 29.

Apart from that, if you have noticed, all the values in this particular data are separated using comma, irrespective of whether we are referring to the column headings or the actual values in the data set every single value is separated using a comma. And that is how data is stored in CSV file. CSV stands for comma separated values. This is one of the most common and popular file type to store most of the data sets.

Now before exploring the usefulness of Pandas, first, let us try to process this file using file handling functions like open and read line. Then only based on the comparison between file

handling approach and Pandas, we will be in the better position to derive a conclusion, which is better. So, let us consider one problem statement, which we have seen earlier many times. What are the total marks of the topper from this scores data set? Let us try to write this Python program using file handling functions.

(Refer Slide Time: 3:28)



Let us look at this code, F is equal to open scores dot CSV in the read mode. We all know, what this open function is then scores is equal to F dot read lines, as we know, this function will read all the lines from the file at a time and after that, we have to do something like 1 colon.

This particular thing is nothing, but our slice operation, which we have done earlier using list, and it is required because we have to skip the first line from the file which holds all those field names because we will not find any marks in that particular line. Hence, the processing will start from second line onwards.

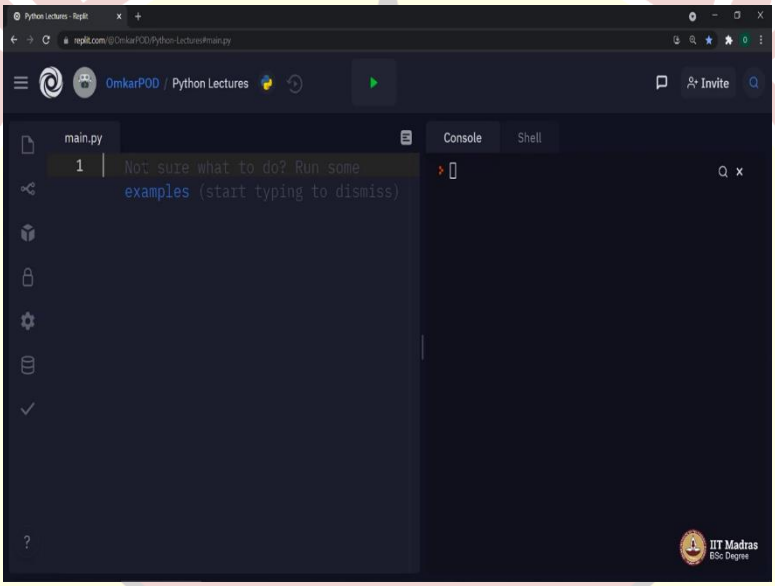
Then max is equal to 0 for every record in scores, which means, the variable record will hold one line from the file at a time. Then we have to split this record using split function over commas in order to get the last value in every single line. Now fields, is a list of all such values. And as there are total 9 fields in the data set, that is why the last field value can be extracted using index 8.

Therefore, int of fields of 8 greater than max. And if that is true, we will update the variable max to int of fields of 8. We will repeat this procedure for every single line in that file. And at the end, we will train the max variable value, which should hold total marks of the topper

student. Let us execute and see. And as expected, he got the result at 281. Now you must be thinking, what is so special about this code.

This is something we have seen with file handling and there is nothing new about it. And if we are supposed to learn Pandas, then why we have not even started with it. And the reason is, look at this particular code. First, it requires 8 lines of code. And on top of that, there are so many different complicated things like this slicing, this splitting then specific index in that particular list a for loop, and so on. Now, we will go to Pandas and try to implement same example in more simplified manner.

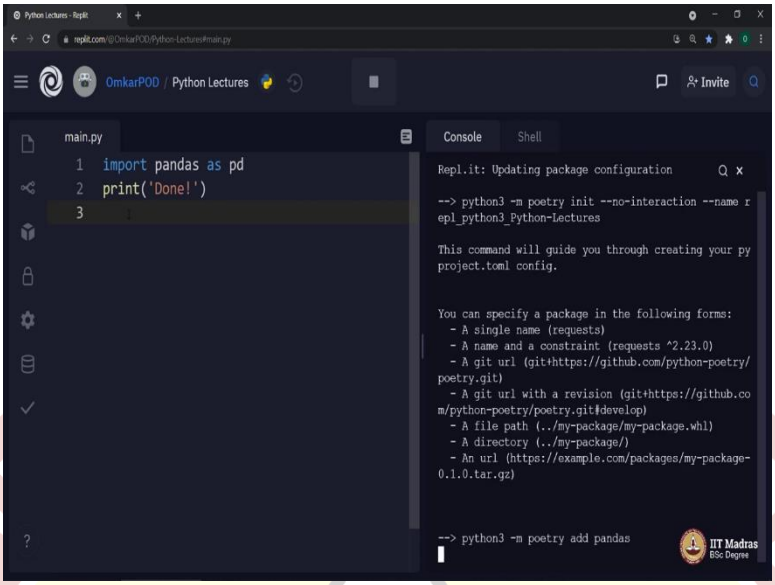
(Refer Slide Time: 6:48)

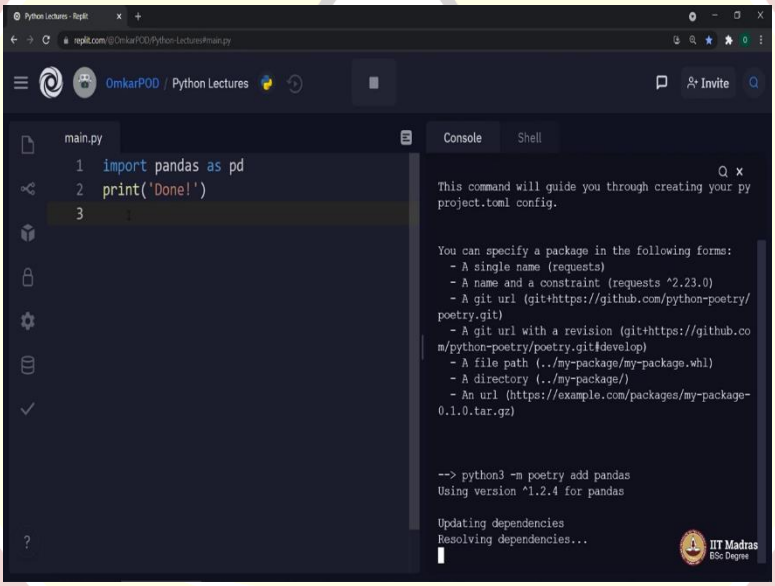


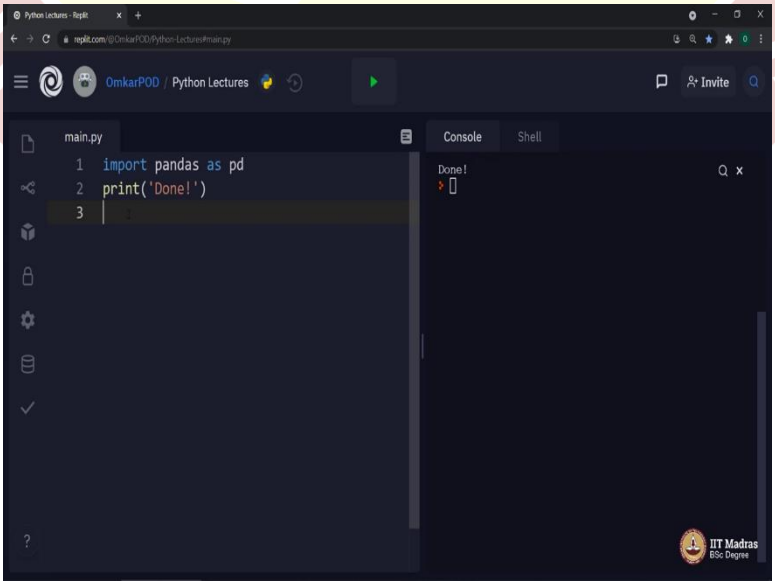
Pandas is also a library hence, we have to import it just like all other libraries, which we have seen earlier like random, calendar, math and so on. But the difference between those libraries and pandas is pandas is an external library, whereas, all those libraries were part of python programming language.

Now what does this mean? This means, Pandas is not part of python language is an external entity, which can be added into python, and then only we can use it. Now you must be thinking how to add Pandas into python? Is it too complicated? Do not worry, there is nothing complicated about it. In fact, you do not even have to do anything additional in order to add Pandas into python. Let us see how it is done.

(Refer Slide Time: 7:56)





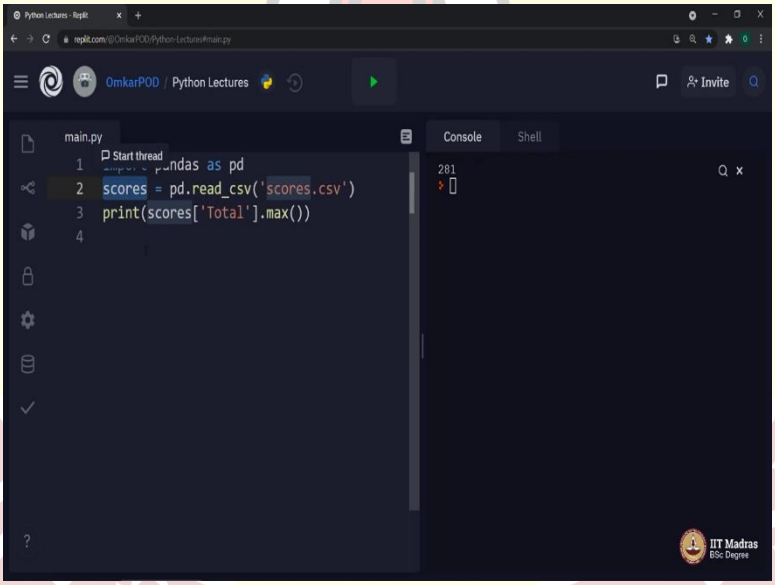


Look at this particular code, import pandas as PD. We all know, this is one of the ways to import a library and then is simply printing one statement, which will indicate that the process is completed. Before I execute this particular code, let me tell you make sure that you are observing the console on the right-hand side of your screen.

Let me execute it. You must have observed python implemented multiple steps in order to install this pandas inside your repl.it and make it available to you. Usually, these installation steps take 10 to 20 seconds depending on your internet speed. Once that is done, now we are all set to use pandas. And as I said, you do not have to worry at all about how to install pandas in Replit.

As you must have seen it is a very straightforward and simple process. Now as the pandas are ready, now let us jump to the program, which we are supposed to execute to find the total marks of the topic students.

(Refer Slide Time: 9:54)



Observe this code. Import pandas as pd. There is nothing new about it, then scores is equal to pd dot read underscore CSV and the name of the file scores dot CSV. This particular function read underscore CSV is used to read a CSV file and store the values into a variable scores. After this, you have written something like scores of total dot max. At this point of time, all these things are new and may even look odd to all of you. But do not worry, we will look into all these details and all these features of pandas in detail in the next lecture.

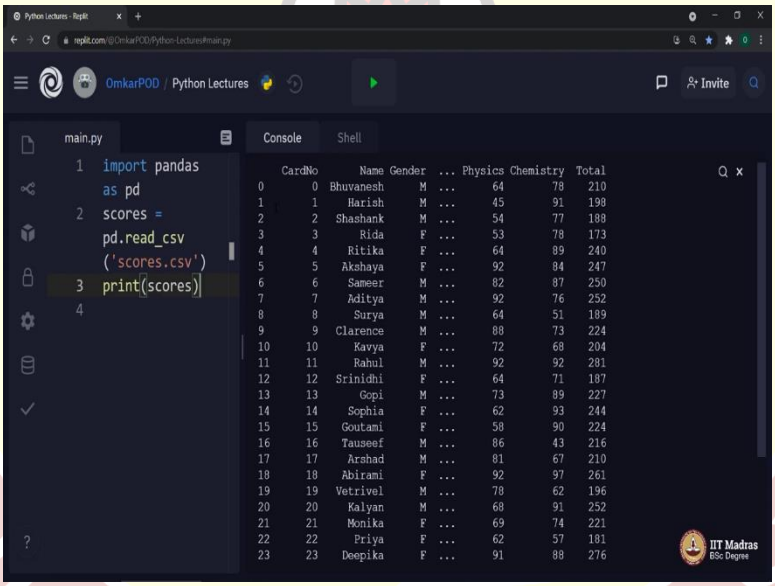
The goal of this lecture is to demonstrate the power of pandas. Hence, we will not go into details of all these things. We will discuss about all these details and various different

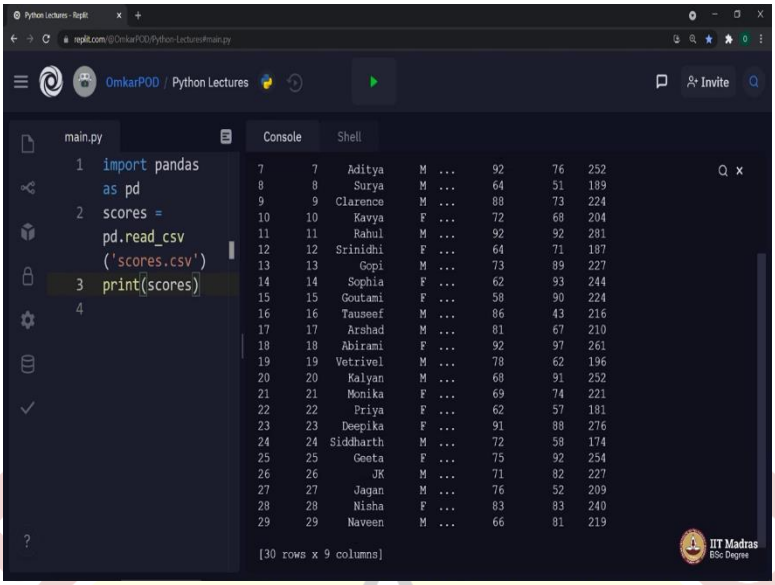
features of pandas in next lecture or so. As of now, do not go in any of these details, which you are currently seeing in these 3 lines of python program. Just observe the output and see how it is easy using pandas to execute some python programs, which were very difficult before this.

Let us execute. As expected, we got output 281. Now if you remember, the code which we wrote earlier using file handling functions it was 8 lines code. Now, we are able to same in 3 lines. And if you observe this very simple and straightforward code, first line for import, second line for read and third line for print, that is it.

Now you must be thinking, is this the only thing we can do using pandas or are there any similar things, which we can do so quickly and so easily. So, let us try a few more things. First, let us see what this course is all about. What exactly it stores.

(Refer Slide Time: 12:36)



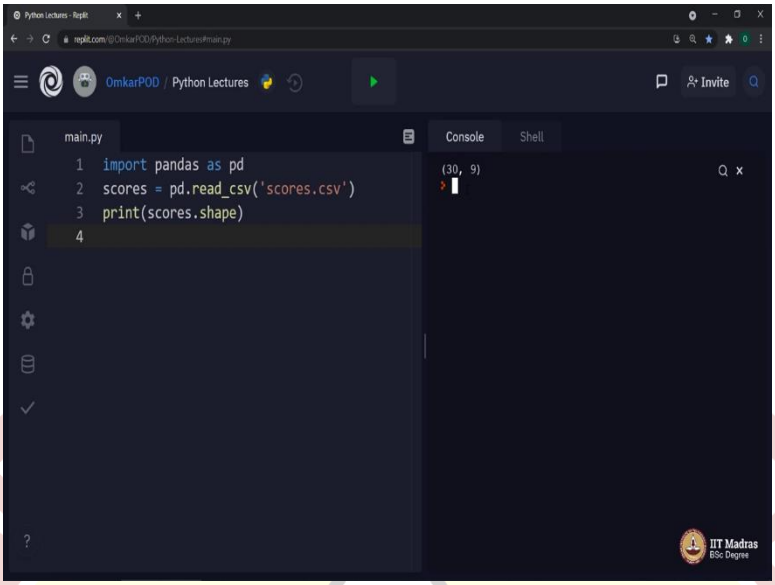


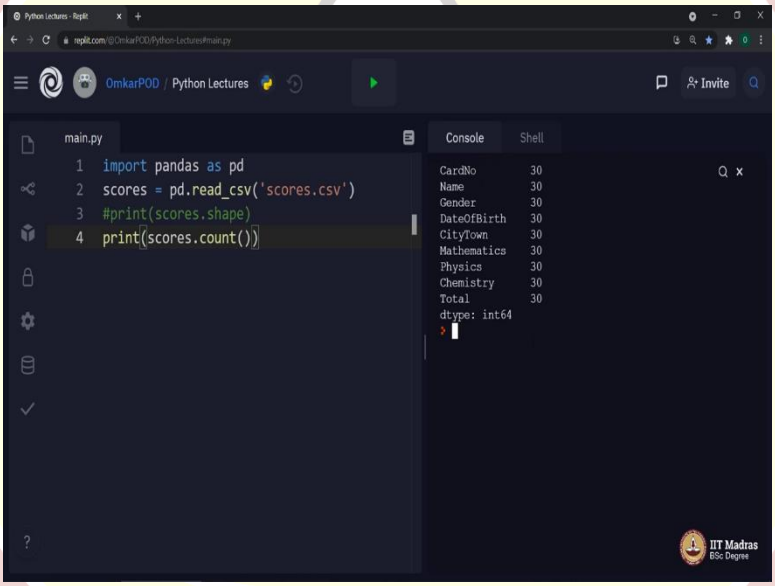
Let us remove all this and print only scores. As you can see, it prints the entire scores data set in the form of a table. Card number from 0 to 29, name, gender, physics marks, chemistry marks and total. If you can see there are some dots in between that means some columns are hidden over here. That is because there is some restriction on how many columns can be displayed at the same time. Even though the columns are not displayed, that does not mean they are removed or anything. They are there only they are not been shown in this particular output.

Apart from that, if you can observe, there are some additional numbers added at the beginning of each line starting from 0 to 29. This is done by Pandas. Pandas always adds a index number for every single line, which you are reading from the file. In our case, we already had card numbers, but still pandas will always add this index numbers, which will be very helpful if you do not have such numbers in your existing file, and at the bottom it says total 30 rows and 9 columns.

Based on this we can easily tell that there are 30 students and 9 different details of every students are there. Like card number name, gender, town, city, date of birth, mathematics marks, physics marks, chemistry marks, and total. Now let us go back to the code and try to execute see more such interesting things.

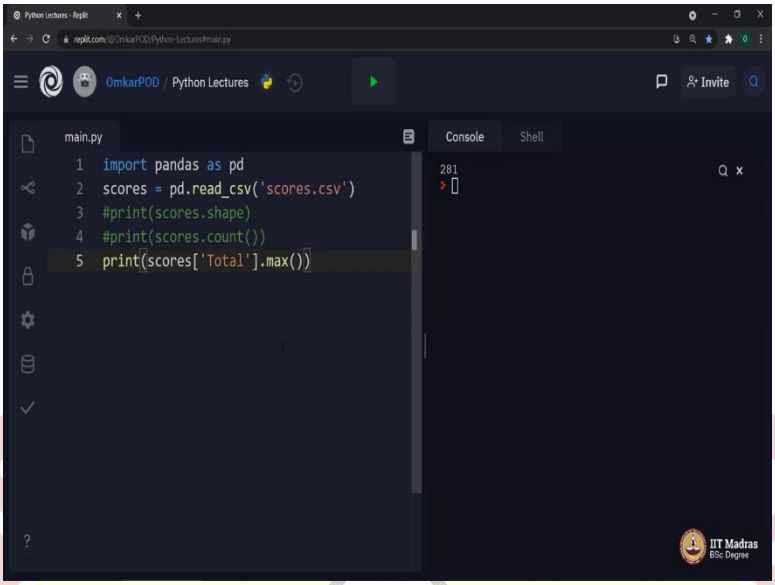
(Refer Slide Time: 14:42)

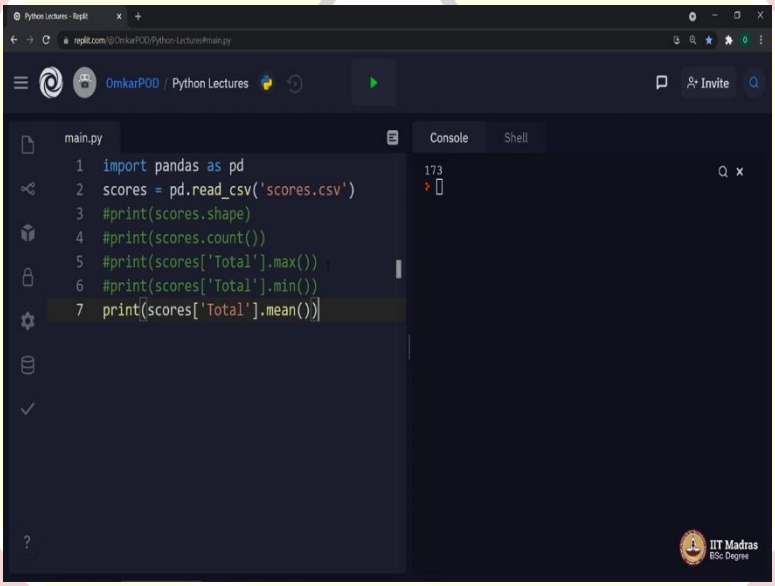




Scores dot shape, 30 comma 9. As you can see, it is enclosed in a round bracket, which means, it is tuple with first value as number of rows and second value as number of columns, and this is one more place where python uses tuples for its internal purpose. Scores dot count. This particular function is displaying how many values are there in each individual column. And as we know, we have values in every column, therefore, for every heading it says 30.

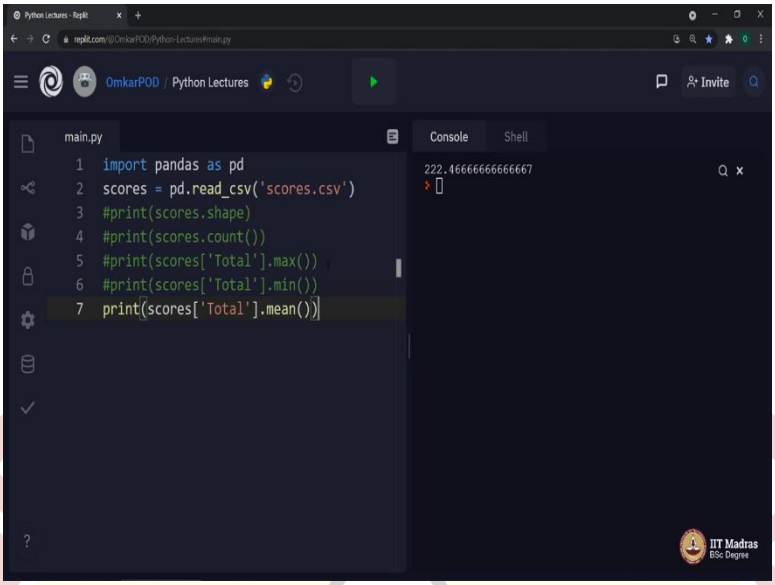
(Refer Slide Time: 15:32)

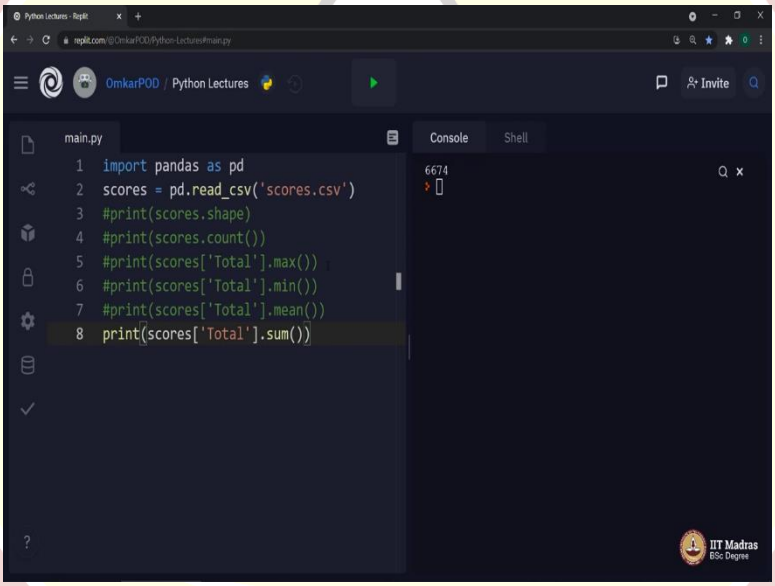




This is max function, which we have already executed. But let us execute it again because I want to show you, it is not just about max, there are so many other things, which you can do in a same manner. Like this mean function, which will give us the minimum value in the total column, which is 173.

(Refer Slide Time: 16:00)



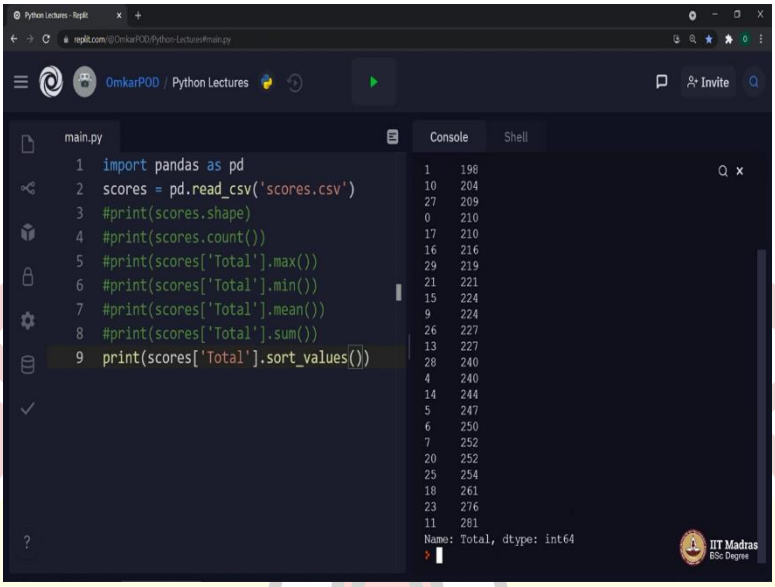


Now let us look at this particular function. This is mathematical mean, which is nothing but average of the column total. Once again, without putting much efforts we are getting average total marks from the scores data set and that too in only three lines, sum, sum of all the values in the total column and we are getting whatever the answer is. Now you must be thinking, is that it or is there anything even more complex, which pandas can do? The answer is, this is just the beginning. Pandas can do so many more things.

And for one more example, let us look at this next line and the function is sort underscore values, which means, we are asking computer to sort all the values in this particular total column. If you remember from earlier, sorting some values required a lot of efforts, especially if we are supposed to write it using some loops, then it even requests nested loops,

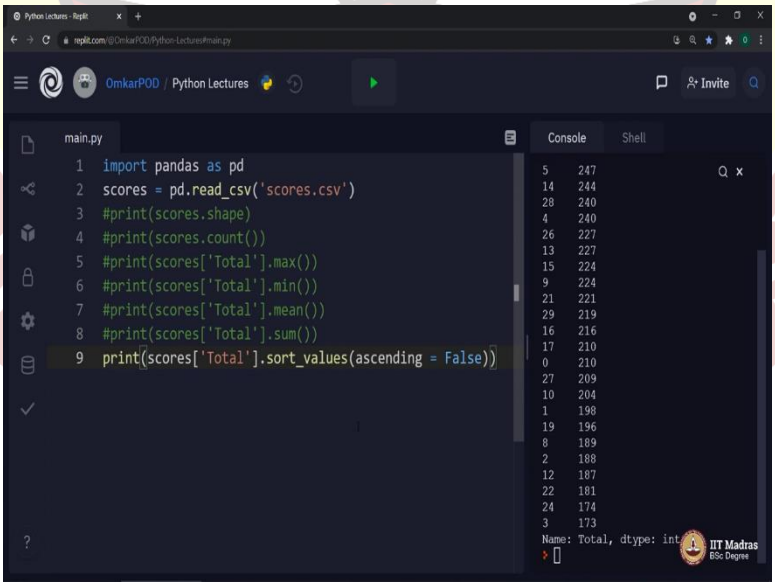
which itself is very complicated. But using pandas, we can do it once again, in just three lines.

(Refer Slide Time: 17:33)



And we have our output from 173 to 281. It is sorting all these values in ascending order, but that is not it, we can also sort these values in descending order just by adding one more parameter to this particular sort underscore values function. Let us see what that parameter is.

(Refer Slide Time: 18:05)



Ascending is equal to false. Let us execute. Now we have values from 281 to 173. As you can see even a complicated code like sorting also can be executed using Pandas without any hassle. Now, just imagine you do not have Pandas and you are supposed to sort the total marks using plain file handling operations. Can you imagine how complex that code will be?

Maybe you should try to write that code, then only you will understand the complexity involved in that particular program.

Now, I hope you all are convinced that pandas is a very powerful tool and using pandas, we can execute so many different complicated codes within very few lines of Python program and that too in a very simpler manner. As I said earlier, the idea behind this lecture was to tell you why we are studying pandas. More details about pandas will be covered in next lecture also. Thank you for watching this lecture. Happy learning