**Data Analytics with Python**
Prof. Ramesh Anbanandam
Department of management studies
Indian Institute of Technology, Roorkee

Lecture No 2
Python fundamentals -1

Good morning students, in the last class, that was the introduction class, we have seen the importance of data analytics and we have seen certain classification of data analytics. This is my second lecture that is Python fundamentals because we are going to use this Python. In this lecture I have 3 objectives.

(Refer Slide Time: 00:50)
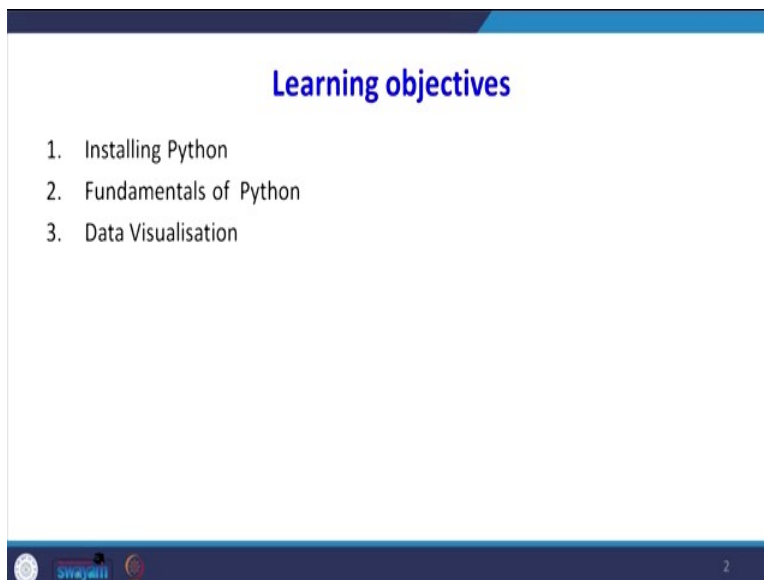
One is I will tell you how to install Python second one I will see some fundamentals of the Python, third one some data visualization. In the data visualization I am going to give only theory in this class. The next class we are going to use Python and we are going to take some sample data and we have to visualize the data using Python software.

(Refer Slide Time: 01:13)

Python Installation Process

Installation Process –

Step 1: Type https://www.anaconda.com at the address bar of web browser.
Step 2: Click on download button
Step 3: Download python 3.7 version for windows OS
Step 4: Double click on file to run the application
Step 5: Follow the instructions until completion of installation process

As I told you the 1st one is how to install this Python. There are 5 steps is there. Step 1, we are going to see in detail in coming slides. In step 1 we are going to visit this website www.anaconda.com at the address bar of the web browser 2nd one we are going to click on download button 3rd one will download python 3.7 version for Windows operating system. Then we will double click that is a 4th step we will double click on file to run the application.

The 5th one will follow the instruction until the completion of installation process. What I have done I have taken the screenshot of all these, all the 5 steps while installing the laptop. I am going to show each steps in the form of screenshot.

(Refer Slide Time: 02:03)

Python Installation Process

Installation Process –

Step 1: Type https://www.anaconda.com at the address bar of web browser.

The 1st one is type this www.anaconda.com at the address bar of the web browser.

(Refer Slide Time: 02:13)

2nd one is once you typed it you can see this screen, here you see this location you can see here. This location there is a download option. When you click that you see that the left side also I have rounded there is a download option you download it.
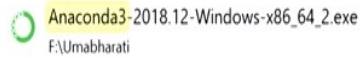
(Refer Slide Time: 02:36)
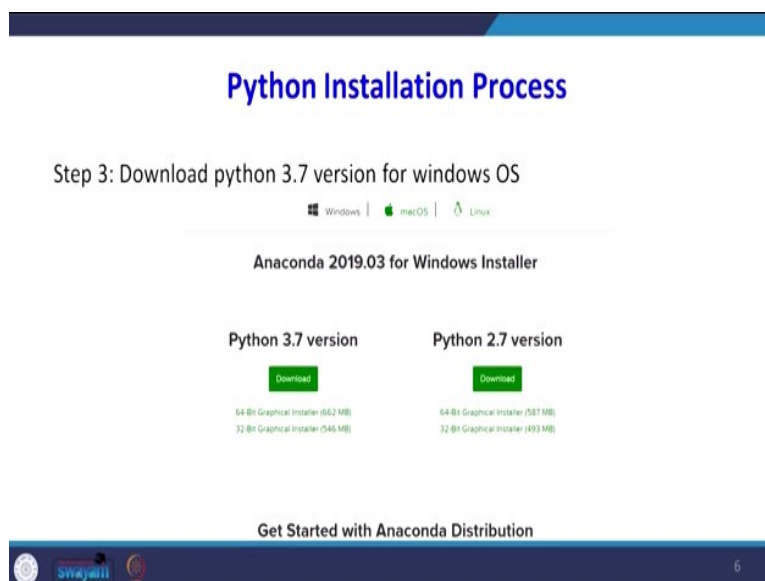
In the 3rd step is there are two versions of python, python 3.7 and python 2.7. In these courses we are going to use the latest version that is the Python 3.7.

(Refer Slide Time: 02:46)

Python Installation Process

Step 4: Double click on file to run the application

Anaconda3-2018.12-Windows-x86_64_2.exe
F:\Umabharati

In the 4th step double click on file to run the application it will get downloaded. when you double click for example; I have stored this anaconda in F folder.
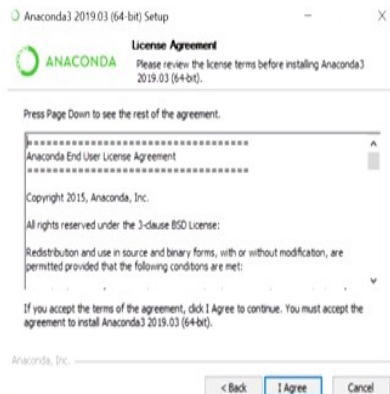
(Refer Slide Time: 02:59)

Python Installation Process

Welcome to Anaconda3 2019.03 (64-bit) Setup

Setup will guide you through the installation of Anaconda3 2019.03 (64-bit).

It is recommended that you close all other applications before starting Setup. This will make it possible to update relevant system files without having to reboot your computer.

Click Next to continue.

Step 5 is just to keep on click Next

(Refer Slide Time: 03:02)

You have to agree for their agreements, terms and conditions.
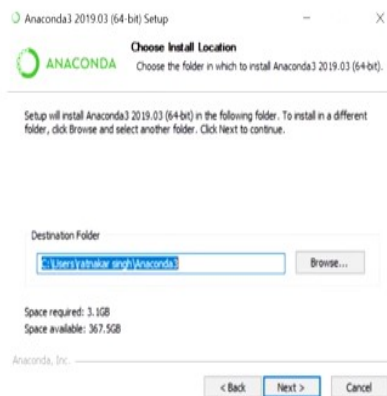
(Refer Slide Time: 03:06)

Next you select, just me recommended click Next.
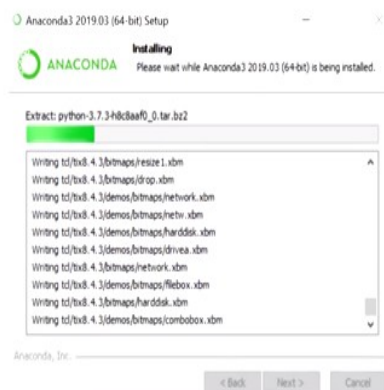
(Refer Slide Time: 03:10)

Python Installation Process

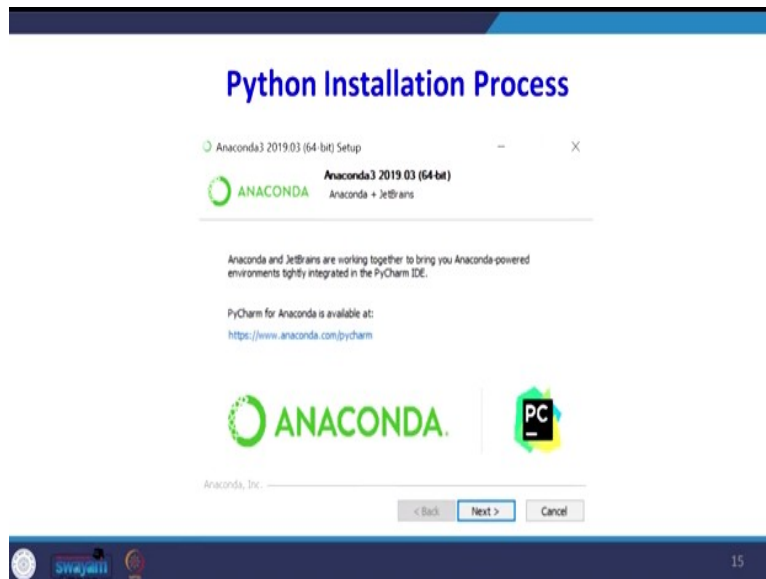Then it is installed in C drive click Next.

(Refer Slide Time: 03:15)

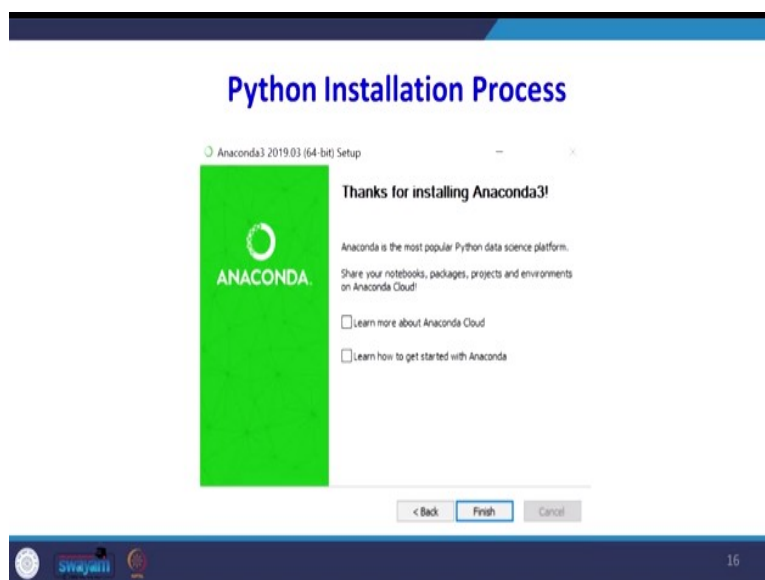**Python Installation Process**

Then install, Installation process is started, then installation is completed.

(Refer Slide Time: 03:25)

Again click Next

(Refer Slide Time: 03:29)
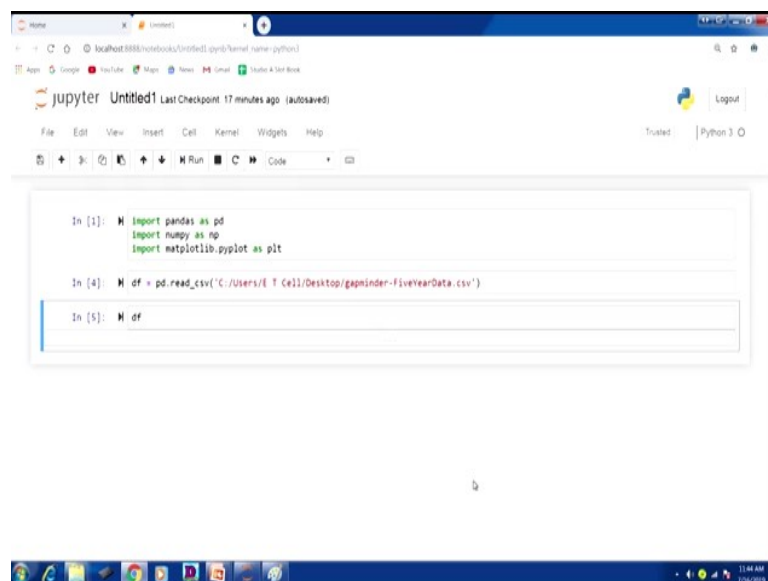

Then click and finish Ok

(Refer Slide Time: 03:34)

Now we have installed the anaconda. So I will explain you how to open Jupyter notebook. I will switch the screen

(Refer Slide Time: 03:46)

Yeah, this is the screen. The initially there are you see, I will see what is this some box it is showing in blue color sometimes it will show in green color that I will show you later. So this is the Jupyter notebook look like.
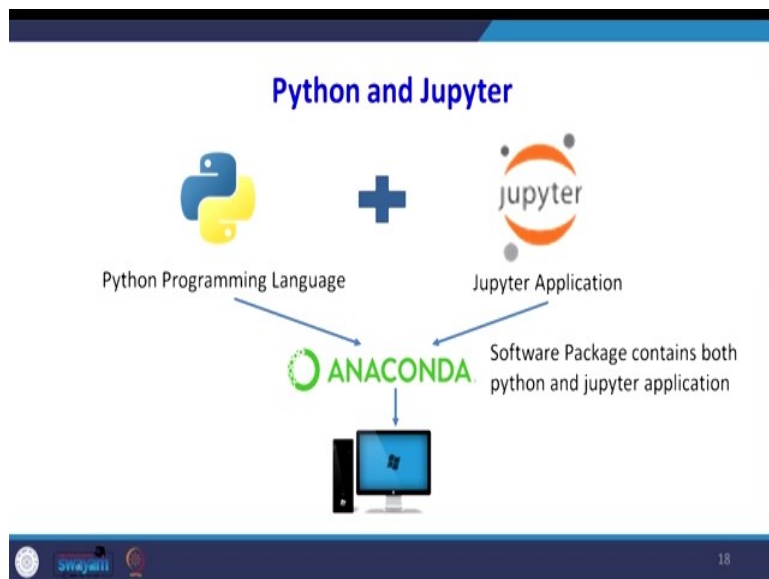
(Refer Slide Time: 04:04)

Why Jupyter NoteBook?

The next one is why there are some more interfaces there for using Python. There is a spider is there Jupyter is it but we prefer Jupyter for some reasons because it is edit code on web browser, it is easy in documentation, it is easy in demonstration and it is user- friendly interface. That was the reason we are using Jupyter it is not necessary if you already you are comfortable in some other interface you can continue with that.
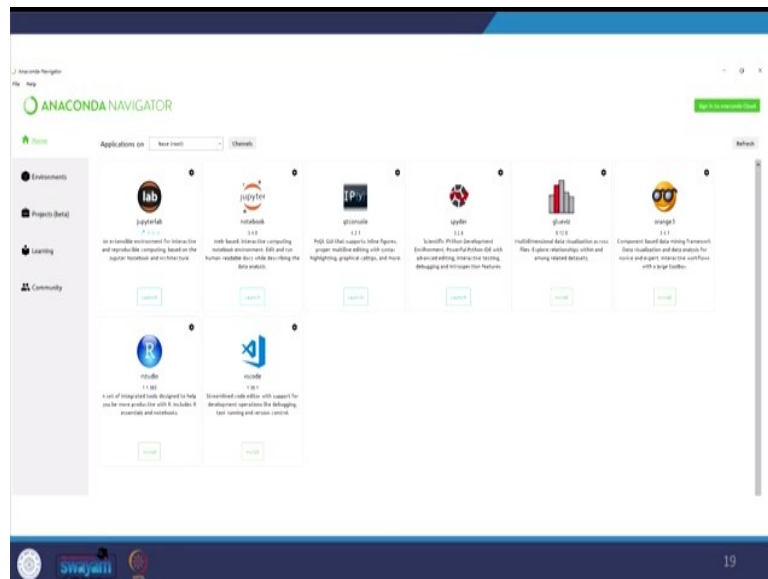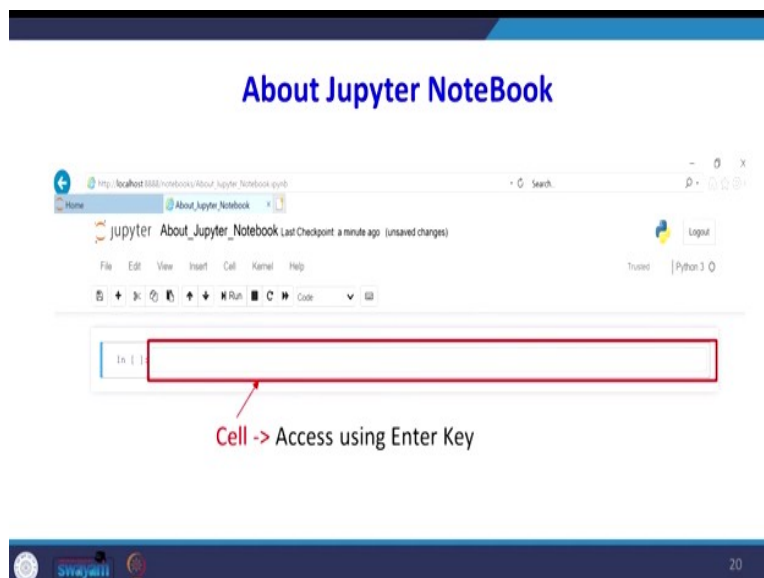
(Refer Slide Time: 04:34)

See that in anaconda it consists of two software one is python that is on the left hand side the another side the right hand side the Jupyter applications these are combined together and kept in the Anaconda software package.

(Refer Slide Time: 04:49)

When you from the start when you type Jupyter you will get this screen

(Refer Slide Time: 04:58)

Then when you click launch you will get this one. So now from the start I am going to explain how to start this Python jupyter notebook.

(Video Starts: 05:08)

You have to type Jupiter and Jupiter notebook .When you click it you will get this one. Suppose if you want to type in a new go new Python 3. Yeah? Here there is a Jupyter there is a it is coming untitled 2 there you can change the name. You give the name as introduction to Python, introduction Okay?

(Video Ends: 05:40)

(Refer Slide Time: 05:40)
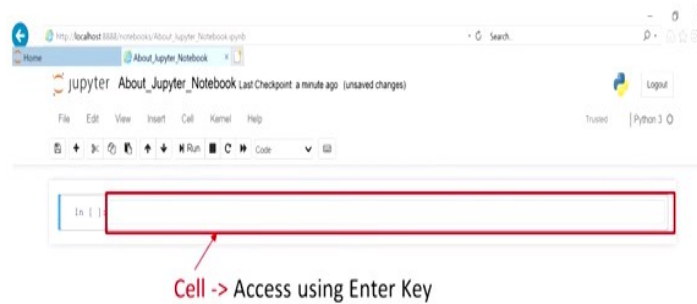
You see there is a box is appearing this is called cell I have made it in the red color, it is a cell it can be the cell can be accessed using Enter key.

(Refer Slide Time: 05:51)

You see sometime that box will look like a green color, Green color indicates it is in edit mode sometime the box will look like in blue color.

(Refer Slide Time: 06:01)

About Jupyter NoteBook
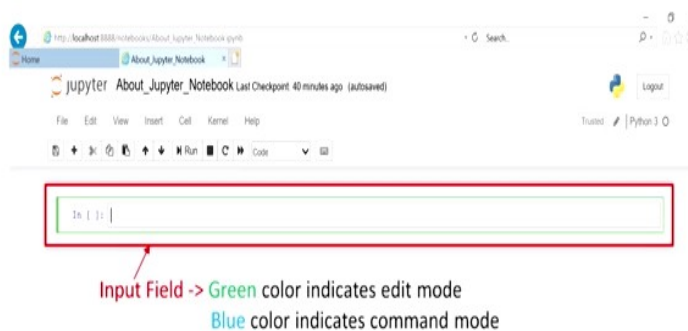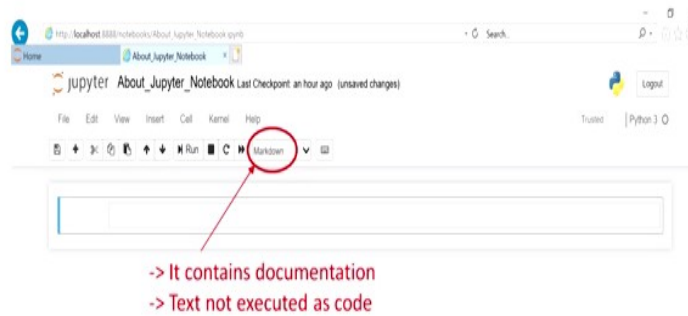
The blue color indicates it is a command mode. See when you go to below the help there is a file name is called mark down .There if you type something then you select mark down that is used for making documentation. So it contains documentation, here text not executed as a code it is only for our understanding purpose.

(Refer Slide Time: 06:22)

About Jupyter Notebook

- Command mode allow to edit notebook as whole
- To close edit mode (Press Escape key)
- Execution (Three ways)
    - Ctrl +Enter (Output field can not be modified)
    - Shift +Enter (Output field is modified)
    - Run button on Jupyter interface
- Comment line is written preceding with # symbol.

Okay? Now about the Jupyter Notebook Command mode allowed editing notebook as a whole. To close edit mode press Escape key. Execution can be done in three ways you can simultaneously we can press Ctrl+Enter. So what will happen when you press Ctrl+Enter output field cannot be modified, another way is to press Shift+Enter output field is modified. Then there is a third way is there is a run button on the jupyter interface.

That you can directly you can click that. Then your code will get executed command line is written proceeding with # tag symbol. So when you want to make some understanding on your program you can use the # symbol, so that will not be executed.

(Refer Slide Time: 07:17)
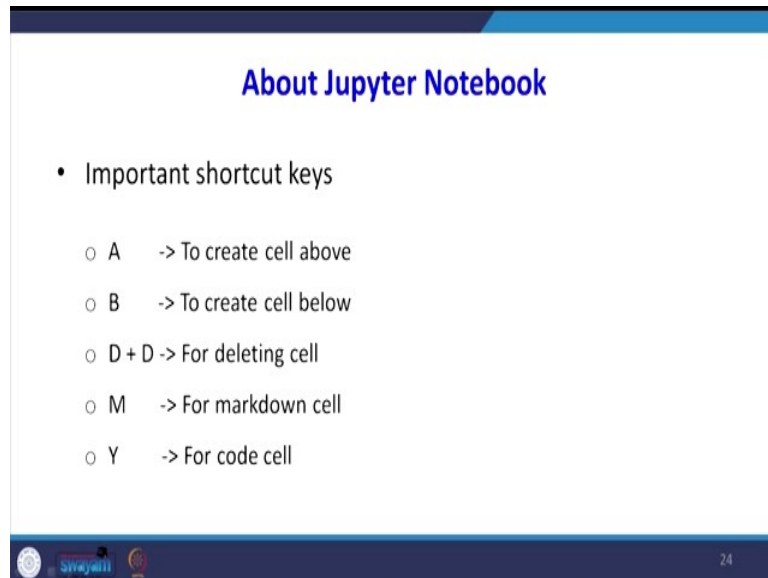
That you only for your understanding purpose but there are about the Jupyter notebook important shortcut keys. When you press A that is used to create a cell above when you press B that is to create a cell below when you press D+ D for deleting cell. When you press M that will made a say mark down cell, when you press Y that is for coding cell.

(Refer Slide Time: 07:46)
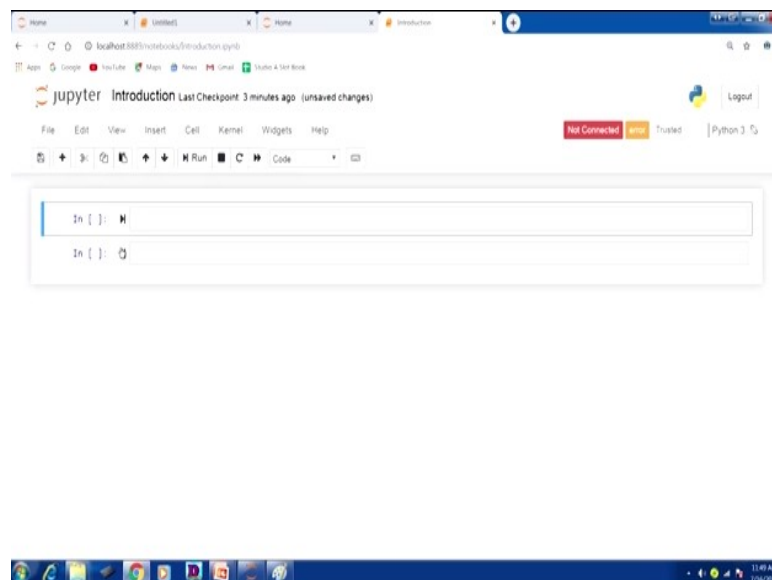
For example; when I am entering B

(Refer Slide Time: 07:54)

Fundamentals of Python

- Loading a simple delimited data file
- Counting how many rows and columns were loaded
- Determining which type of data was loaded
- Looking at different parts of the data by subsetting rows and columns

We will go to the next one fundamentals of Python and you see loading here .What we are going to see in coming slides. Loading a simple delimited data file counting how many rows and columns were loaded and determining which type of data was loaded. Then looking at different parts of data by subsetting rows and columns because these activities are more important because once we loaded a data that may have n number of cells n number of rows, column and rows.

Sometime we need to do some operation using only few rows are few cells .You should know how from a big data file how to use only a particular row or how to use a particular column. Sometimes we can have a collection of rows also, collection of columns also for doing our specific operations.

(Refer Slide Time: 08:49)

Pandas for Everyone
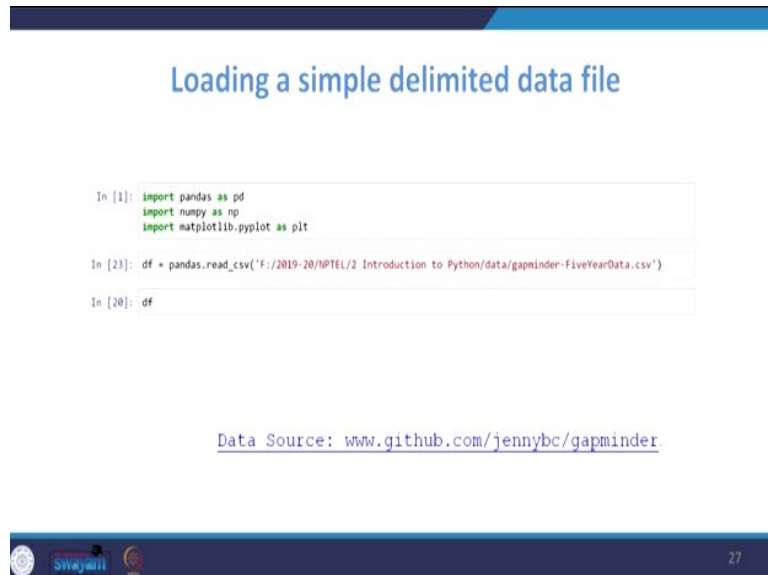
Python Data Analysis

Daniel Y. Chen

♦ Addison-Wesley

Boston • Columbus • Indianapolis • New York • San Francisco • Amster-dam • Cape Town
Dubai • London • Madrid • Milan • Munich • Paris • Montreal • Toronto •
Delhi • Mexico City
São Paulo • Sydney • Hong Kong • Seoul • Singapore • Taipei • Tokyo

This was the reference book which I am following for this course and the book name is Pandas for everyone especially for this lecture. It is the professor Daniel Y. Chen he is the author of this book.

(Refer Slide Time: 09:04)

Loading a simple delimited data file

```
In [1]: import pandas as pd
        import numpy as np
        import matplotlib.pyplot as plt

In [23]: df = pandas.read_csv('F:/2019-20/NPTEL/2 Introduction to Python/data/gapminder-FiveYearData.csv')

In [20]: df
```

Data Source: www.github.com/jennybc/gapminder.

Now we are going to learn how to load a simple delimited data file. This is the fundamental because before doing data analysis the first step is how to load the data into the Python. For that purpose we are going to import some basic libraries one is pandas numpy another is matplotlib.pyplot as plt. So, first we are going to import these three basic library .Then we are going to load the data. The data sources it is taken from www.github.com/gennybc/gapminder.

So I have downloaded this data set already I am going to tell you how to load the data set in to the python. Before that I am going to open that excel file, I am going to show what is the column? What is the row open the excel file?

(Video Starts 10:07)

When you look at this I am reading the column see that there is a country, year, population, continent, life expectancy that is given as the short name life exp then gdp per capita. So in rows there are, how many rows is there I will tell you how many rows is there I am coming down and this is a this is csv file format. How many rows are there? There are 1705.

The last row is Zimbabwe Right? Please look at the data Zimbabwe year 2007. I think it is a population, continent, life expectancy this is a per capita income. Okay? Now this data, this csv file I am going to import into the Python. You see that I am going to call this data set df.

df= pd because pd is the short-form of pandas, Pandas nothing but the panel data Pandas.read_csv.

Why I am using csv because the csv file is I am going to read it. The location of the file given the path of that file you can directly copy that path but one thing you have to note it down. See, C: this will be this should be \ because when you copy that path directly. Generally you will get here / but you have to change it. So I changed it back C: / users / ET cell / desktop / gapminder-five year data.csv.

Look at their it should be in the code. Now I am going to read the df, Yes? once I read it you see that, the row is starting from 0 . That is a very important. It is a 0 indexing 0, 1, 2, 3, 4, 5, 6, 7, 8 I am able to see whatever I have seen in the csv file just a few minutes before. You see I am able to see the country, year, population, continent, life expectancy and gdp per capita. Okay? What I how I have read it pd. read_csv

Suppose I have installed, I have loaded that data I want to say what are the headings of that file. Heading means what are the columns. For that there is, in Python there is a two type print and open the parentheses df.head when you execute this one you will get 1st 5 rows that means 0, 1, 2, 3, 4. So that means when you execute this one you can see 1st 5 rows from the data set Yes? You are able to see that, Okay?

I will go to the next command, suppose I want to know the size of that file that is I want to know how many rows and how many column is there. For that there is a command called the shape. So print df.shape, df is they were finally because we outer loading that csv file we have named in the variable called df. Okay? So when you type print df.shape then we will come to know how many rows are there. How many columns are there?

So, I am typing print df shape. One more thing you should not type this parenthesis because it is the shape is without parentheses. So I am going to remove this parenthesis again I got to run it. Yes, it is showing how many rows? How many columns? Okay? We will go to the next one now I want to know how many column names? What are the column names? So if I type print df.columns Right?

Here, please note that here also there is no parenthesis if I type print df.columns. This was the output which I copied see what is output disappearing country, year, population, continent, life expectancy, gdp per capita, data type is object; I will show you how this comment is running. Type print df., yes you are able to get this way. So what the students what you have to do while looking at the video you have to open your laptop you have to type this command.

Then you have to see you can verify the answer. Okay? The next command is to get the data type of each column; you have to type this command print df.dtypes. That will give you the summary of the all data set and what is the nature of the data. We will see that how it is appearing. So, I am going to type print df.dtypes. Now you we will see the data type of each column. For that you have to use this command dtypes.

So print df.dtypes, this is the output which you will get it. I will show you in the Python, first we will look at what is the subject output you see countries object, year is an integer, population pop it is a variable that is in the float. Float means there is a decimal continent it is an object that means a character, life expectancy it is a float that means you are going to get that value in decimals.

Similarly, gdp per cap that also going to get in the decimals then data type is object now we will go to the Jupiter. We run this command so you see that you see line number 8. Print df. dtypes you are getting whatever it was there in this or whatever I have shown in the slide is there. Say country object, year integer, population float type of data and so on.

(Video Ends 17:10)
(Refer Slide Time: 17:11)

Pandas Types Versus Python Types

Pandas Type	Python Type	Description
object	string	Most common data type
int64	int	Whole numbers
float64	float	Numbers with decimals
datetime64	datetime	datetime is found in the Python standard library (i.e., it is not loaded by default and needs to be imported)

This is a classification of types of data in the perspective of pandas, in the perspective of Python. See when they say string it is a most common data type it is a character. When I say 'Int' it is a whole number integers. It is a float number with the decimals. Date, time is that is to represent the data it is not loaded by default that need to be imported. Whenever it is a requirement is there that we will see

(Refer Slide Time: 17:38)

get more information about data

```
print(df.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1704 entries, 0 to 1703
Data columns (total 6 columns):
country     1704 non-null object
continent   1704 non-null object
year        1704 non-null int64
lifeExp     1704 non-null float64
pop         1704 non-null int64
gdpPercap   1704 non-null float64
dtypes: float64(2), int64(2), object(2)
memory usage: 80.0+ KB
None
```

That one more command is to get more information about the data. So you type df.info you will get the full details about each columns. We will do that one.

(Video starts: 17:53)

Look at this when it print df. info so I am getting data columns there were 6 columns country there are 1704 rows is there Non null object that means all the data is there is no missing values. Similarly year 1704 rows is there, Non-null that is an integer, Non null means, that all

the values are filled. There is no missing cell so population, float, continent object, life exp float, gdp per capita float, memory usage is this much.

Suppose, there is a big data file is there we want to see the specific rows are specific columns. How to do that? Now get the country column and save it to its own variable. So country if you look at the data which I will show initially countries one of the column. So I want to pick up only that country column I am going to save it. I am going to give the name for that a country_df= df you see that you have to open the square bracket, Square bracket within quote.

Suppose, in the country column I want to see 1st 5 rows Okay? You type print, open parenthesis country_df.head that shows 1st 5 rows and see that now from the full data we have fetched only the country column. That we have seen there are 1st 5 rows, that is 0th row is a, 0, 1, 2, 3, 4, 5 to 5 rows we are able to see when you from the big file. Suppose there may be requirement you need to see what are the last five observations for that purpose.

You type print country_df.tail, then you can see from the bottom we can see last 5 rows. You will see how it is appearing, Yes? So what is it we are able to see last 5 rows from the country, country_df file.

(Video ends: 21:15)

(Refer Slide Time: 21:15)

There may be requirement you need to see more than one column at a time. So I am going to save in the form of another file name that is called a subset, Subset=df. You see there is a double square bracket so I want to switch the country columns, continent columns and year columns. Then I going to see what are the heading that means I want to see what does the 5 rows of these subsets so we will go to they go to Python.

(Video starts: 17:46)

I am going to call it a subset continent. Suppose I want to see the 1st 5 rows of this file called a subset. Data set called subset. You see that I am able to fetch 3 columns at a time that is on the country, continent and year. The same way we from the subsets file I want to see last 5 rows so print subset.tail. Let us see what we are getting we will get this output Yes? You see that there are 3 columns.

There were the last 5 rows from the button. So far we looking at different columns now we want to subset rows by index label there is one command called loc. So first we look at the file initial file that is a print df.head. Next you see that I want to locate the 0th row so for that purpose, print df.loc see it is a square bracket you type 0 because if you suppose we want to know the 1st row i out to enter because I would enter 0 because Python counts from 0, so print df.loc 0 that will show the 1st row.

You see 0th row access country Afghanistan year 1952 population is this much continent is Asia. Suppose I want to access this 1st row that means 0th row, Yeah? 0th row you can verify 0th row is the country Afghanistan, year 1952 this is a way to access a particular row. Dear students whatever comments which I am typing that I will be given to you when you take this course you can practice yourself.

You need not bother about in case we are not getting at this stage this all the commands all the codes will be given to you .You can practice on yourself. Suppose I want to get the 100th row how to access from the file df? I want to look at 100th row so you type print df.loc 99. You can exactly access in 100th row what is the element is there? Suppose I want to access 100th row df, 100th row is the country Bangladesh, year 1967, population this is.

This is the way to access different rows for our calculation purpose. So far we have seen how to load csv file into the Python, we have seen some basic commands.

(Video Ends: 25:21)

We have seen how to know the size of the file then we have seen how to access a particular row and also we have seen how to subset from the given big file? How to subset different small data file? So that can be used for our further analysis. So the next class we will see how to access different columns that will continue in the next lecture. Thank you.