



APSSDC

Andhra Pradesh State Skill Development Corporation



Data Analysis Using Python



NumPy

pandas

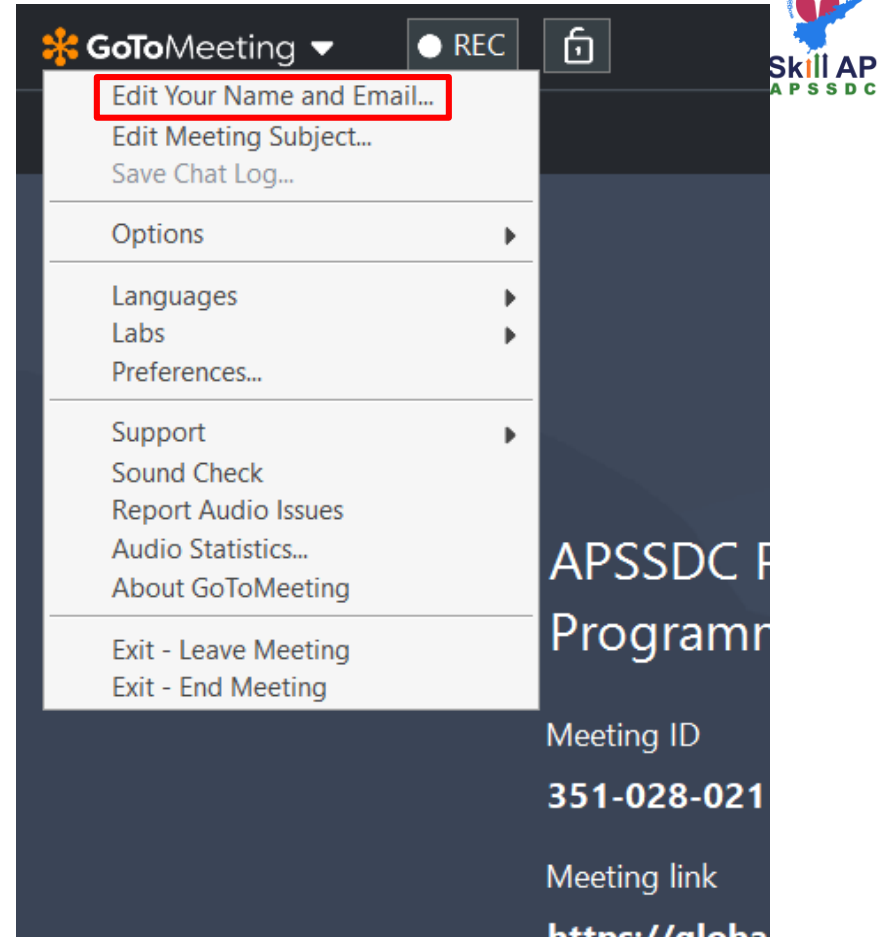


matplotlib



For Attendance and
Verification Purpose

**RollNo-Name
And Registermed
Email ID**



Session Resources

<http://bit.ly/apssdc-da-eb5>

Why you are attending this training

To gain some knowledge

To learn

To apply it

To improve our skills

As data analysis was trending technology to implement ai,ml applications

Steps involved in Data Analysis

1. Understand the data
2. Analyze the data
3. Clean the data
4. Preprocess the data
5. Visualize the data

Tools/Programming Languages for Data Analysis

R

Python – Easy to Learn, Huge Community, open source, os packages

Matlab

Excel

SQL

Power BI, Tableau,

What are the trending technologies

1. Quantum Computing
2. AI
3. ML
4. IoT
5. Data Science
6. Block Chain
7. Augmented Reality and Virtual Reality
8. Cognitive Cloud Computing
9. Angular and React
10. DevOps
11. Internet of Things (IoT)
12. Big Data
13. RPA (Robotic Process Automation)

First 4 Days Agenda

Introduction to
Data and Data
Analysis Using
Python

Conditional
Statements
Loop and Data
Structures

Programming
using Python

File, Packages
and Functional
Programming

Agenda Next 8 Days

Intro to Data
and Data
Manipulation
with NumPy

Data Analysis
with pandas

Data
Preprocessing
with Scikit-
Learn

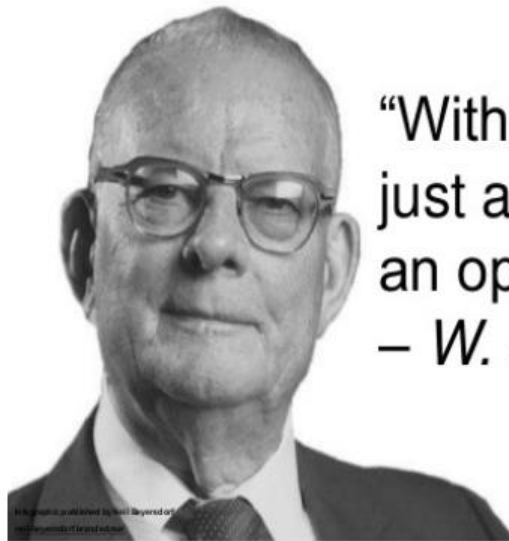
Cleaning Data
in Python

Introduction
to Data
Visualization
& Matplotlib

Data
Visualization
using
Seaborn

What is Data?

Data are facts and statistics collected together for reference or analysis.



“Without data you’re just another person with an opinion.”

– *W. Edwards Deming*

Data Collection

Primary Data Sources - Data Collected at Source

Examples are

Surveys, Design of Experiments, Simulation, Web Scraping, Social Media Extraction, IoT Sensors, etc. Note: IoT sensors data can also be secondary data.

Pro: Get data for exact variable of interest

Con: Costly & Time Consuming

2. Secondary Data Sources - Data Collected Before Hand

E.g. RDBMS, CRM, FRM, SCM, HCM, SQL Databases, etc. Also publicly available data Open Source as well as Syndicate Data

Pro: Data is easily available

Con: Data may or may not have variable of interest

Interesting insights

Bombardier showcased its C Series jetliner that carries Pratt & Whitney's Geared Turbo Fan (GTF) engine, which is fitted with 5,000 sensors that generate up to 10 GB of data per second. A single twin-engine aircraft with an average 12-hr. flight-time can produce up to 844 TB of data.

Saudi Aramco laid 650km of new pipelines across a mountain range of red sand dunes. How do they monitor that?

Using 100,000 sensors and data points on wells, pipelines, plants and terminals, it directs every drop of oil and cubic foot of gas that comes out of the kingdom

One study predicts that by 2020, 1.7 MB of data will be created every second for every person on earth.

The average number of AI projects for a business is expected to increase to 35 by 2022 from four this year, according to a Gartner Inc. survey of about 100 organizations of various sizes, many of them with annual revenue of \$1 billion to \$3 billion. The research and advisory firm also said the number of its clients requesting help in dealing with AI suppliers grew 57% between 2017 and 2018.

As per the report by NASSCOM and Blueocean, India is reigning big data analytics with a value of \$1.2 billion placing it among the top 10 big data analytics markets in the world. They have also anticipated the growth becoming eight-fold by 2025, soaring to \$16 billion. With this vision in mind, every sector is now looking forward to Data analytics for its evolution.

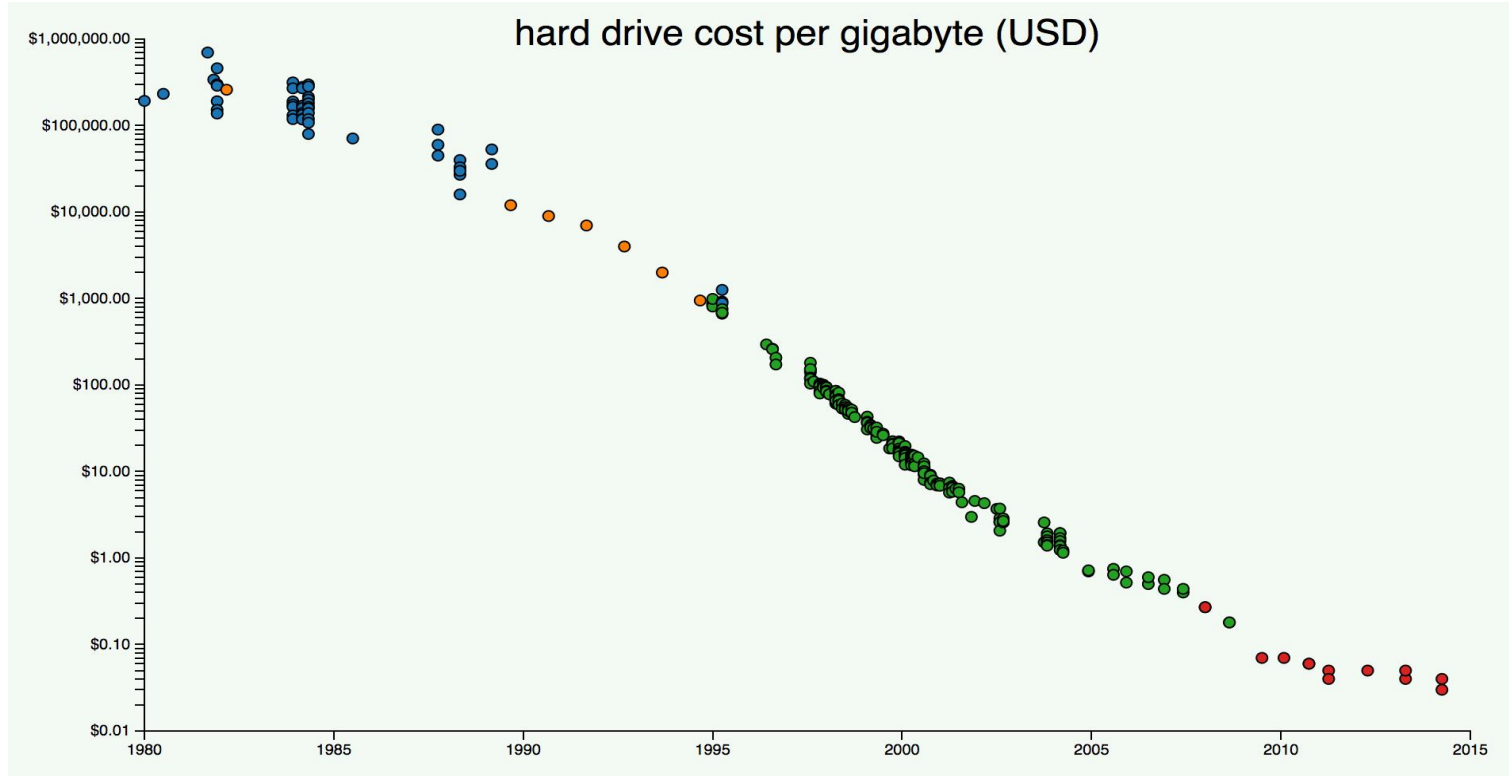
Storage capacity, size



512GB



cost



Data Generation

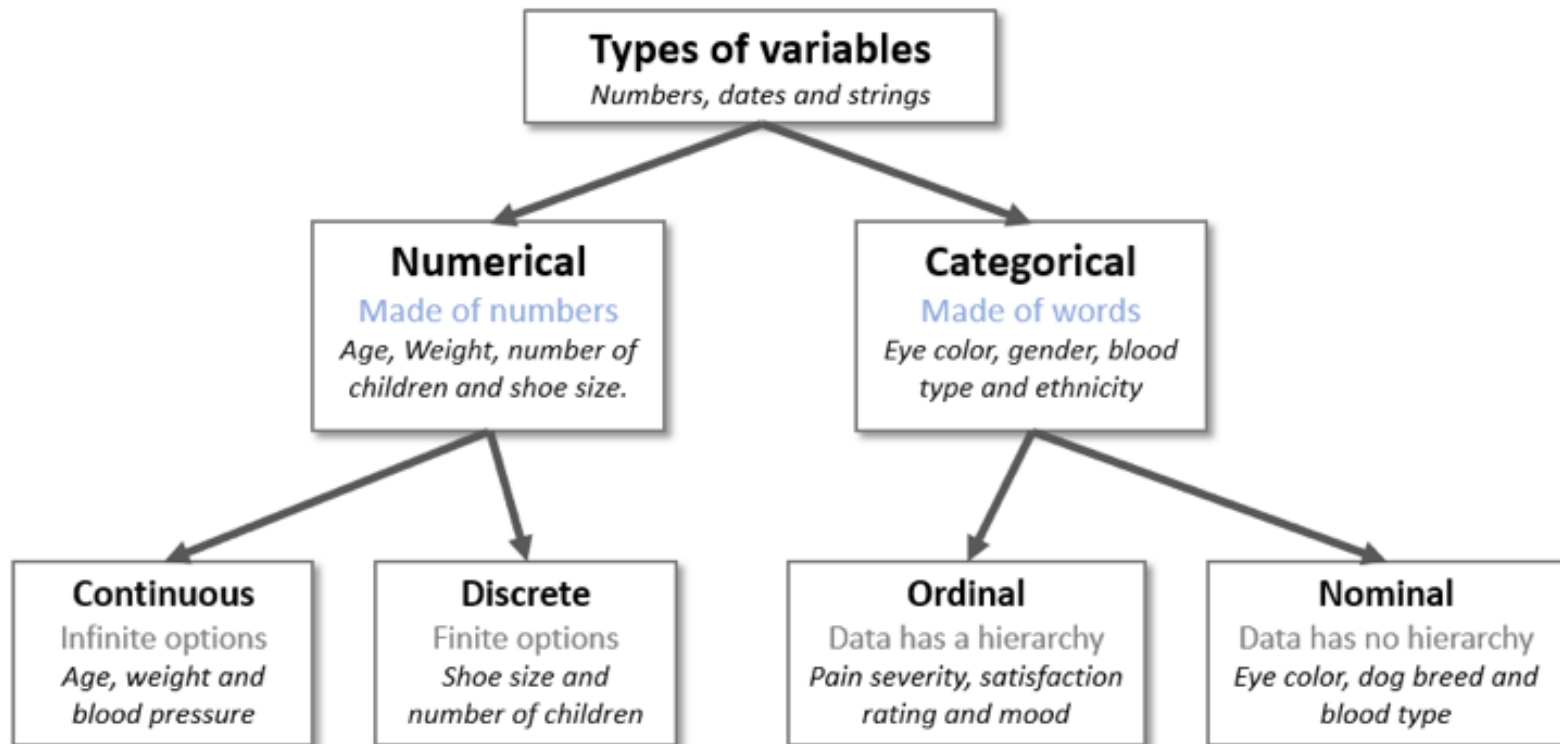


Types of Data

1. Structured Data – row/columns → Tabular format → Excel, .csv, .tsv,
2. Unstructured - - Pics, videos, text, ppt, pdf, logFiles, voices
3. Semi-structured → xml, json, html files

{key:value}

DATA TYPES IN STATISTICS



NUMERICAL DATA

1. DISCRETE DATA

If its values are distinct and separate. In other words: We speak of discrete data if the data can only take on certain values. This type of data can't be measured but it can be counted. It basically represents information that can be categorized into a classification. An example is the number of heads in 100 coin flips.

You can check by asking the following two questions whether you are dealing with discrete data or not: Can you count it and can it be divided up into smaller and smaller parts?

2. CONTINUOUS DATA

Continuous Data represents measurements and therefore their values can't be counted but they can be measured. An example would be the height of a person, which you can describe by using intervals on the real number line.

Contd..

Interval Data

Interval values represent ordered units that have the same difference. Therefore we speak of interval data when we have a variable that contains numeric values that are ordered and where we know the exact differences between the values. An example would be a feature that contains temperature of a given place like you can see

Temperature?

- ☐ - 10
- ☐ - 5
- ☐ 0
- ☐ + 5
- ☐ + 10
- ☐ + 15

CATEGORICAL DATA

Categorical data represents characteristics. Therefore it can represent things like a person's gender, language etc. Categorical data can also take on numerical values (Example: 1 for female and 0 for male). Note that those numbers don't have mathematical meaning.

NOMINAL DATA

Nominal values represent discrete units and are used to label variables, that have no quantitative value. Just think of them as labels. Note that nominal data that has no order. Therefore if you would change the order of its values, the meaning would not change. You can see two examples of nominal features in the right.

The left feature that describes a persons gender would be called „dichotomous“, which is a type of nominal scales that contains only two categories.

What is your Gender?

- ☐ Female
- ☐ Male

What languages do you speak?

- ☐ Englisch
- ☐ French
- ☐ German
- ☐ Spanish

Contd..

ORDINAL DATA

Ordinal values represent discrete and ordered units. It is therefore nearly the same as nominal data, except that its ordering matters. You can see an example below:

What Is Your Educational Background?

- ☐ 1 - Elementary
- ☐ 2 - High School
- ☐ 3 - Undergraduate
- ☐ 4 - Graduate

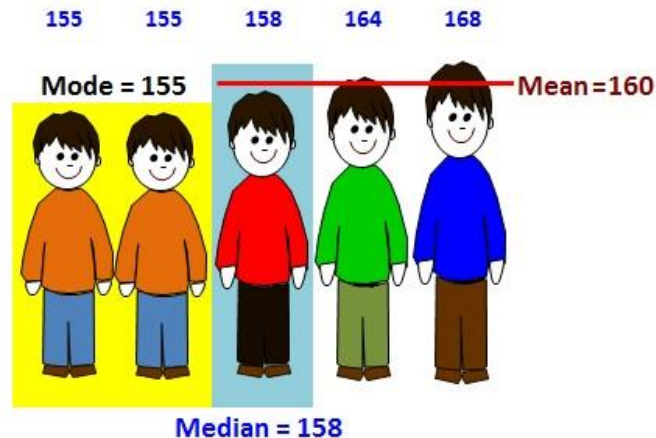
Note that the difference between Elementary and High School is different than the difference between High School and College. This is the main limitation of ordinal data, the differences between the values is not really known.

What is Statistics?

A branch of mathematics that takes and transform the data into some useful information which in turn is used to make some decisions.

Statistics is concerned with

- Processing and analyzing data
- Collecting, presenting and transforming data to assist decision maker



Measures of Dispersion

Range: It is the difference between highest value and the lowest value in the data set.

For a given list of numbers: 10, 20, 40, 10, 70 the range is $70 - 10 = 60$.

Variance: The average of the squared differences from the mean.

Steps to calculate variance:

- Calculate mean (mean is nothing but average)
- Find difference of each data from mean
- Square all the differences
- Take the average of the squares.

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Standard Deviation: It shows you how much your data is spread out around the mean. Its symbol is σ (the Greek letter sigma). It is the square root of the **variance**.

Why python?

Why python



have caused uncertainties on the country's job market, but the demand for niche job skills has not dried up. Corporates are hiring talent to select roles that require specialised certifications in niche skills across business segments.



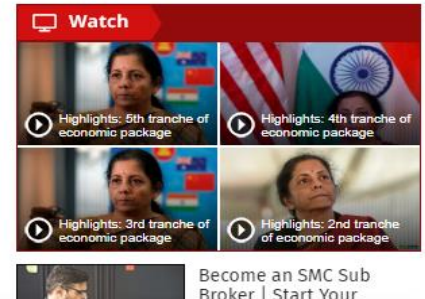
Moneycontrol gives you a lowdown on the top 10 skills in demand across Indian corporates in times of COVID-19:

Python programming language

Python is the second most loved programming language, according to StackOverflow developer survey, and for a reason. It is easier to learn, efficient and is usually the programming language taught in schools and colleges.

RELATED NEWS

So it is one of the most preferred languages for data scientists, Artificial



Features of Python

Easy To Learn,
Code And Read

Free And Open-
source

High-level
Programming
Language

Portable And
Extensible

Interpreted

Object-oriented

Embeddable

Large Range Of
Library

GUI Programming

Dynamically
Typed

PYTHON PROGRAMMING APPLICATIONS

Python had been developed to assimilate and work dynamically across various platforms. Here is a list of applications on its functional role:

1. Artificial
Intelligence

2. Machine
Learning

3. Data Analysis

4. Web
Development

5. Game
Development

6. Embedded
Applications

7. Scripting
Applications

History

Python is an interpreted, high-level, general-purpose programming language.

- 1994 -----> v1.0
- 2000 -----> v2.0 – 2020 2.8
- 2008 -----> v3.0
- 2019 -----> v3.8 3.7+



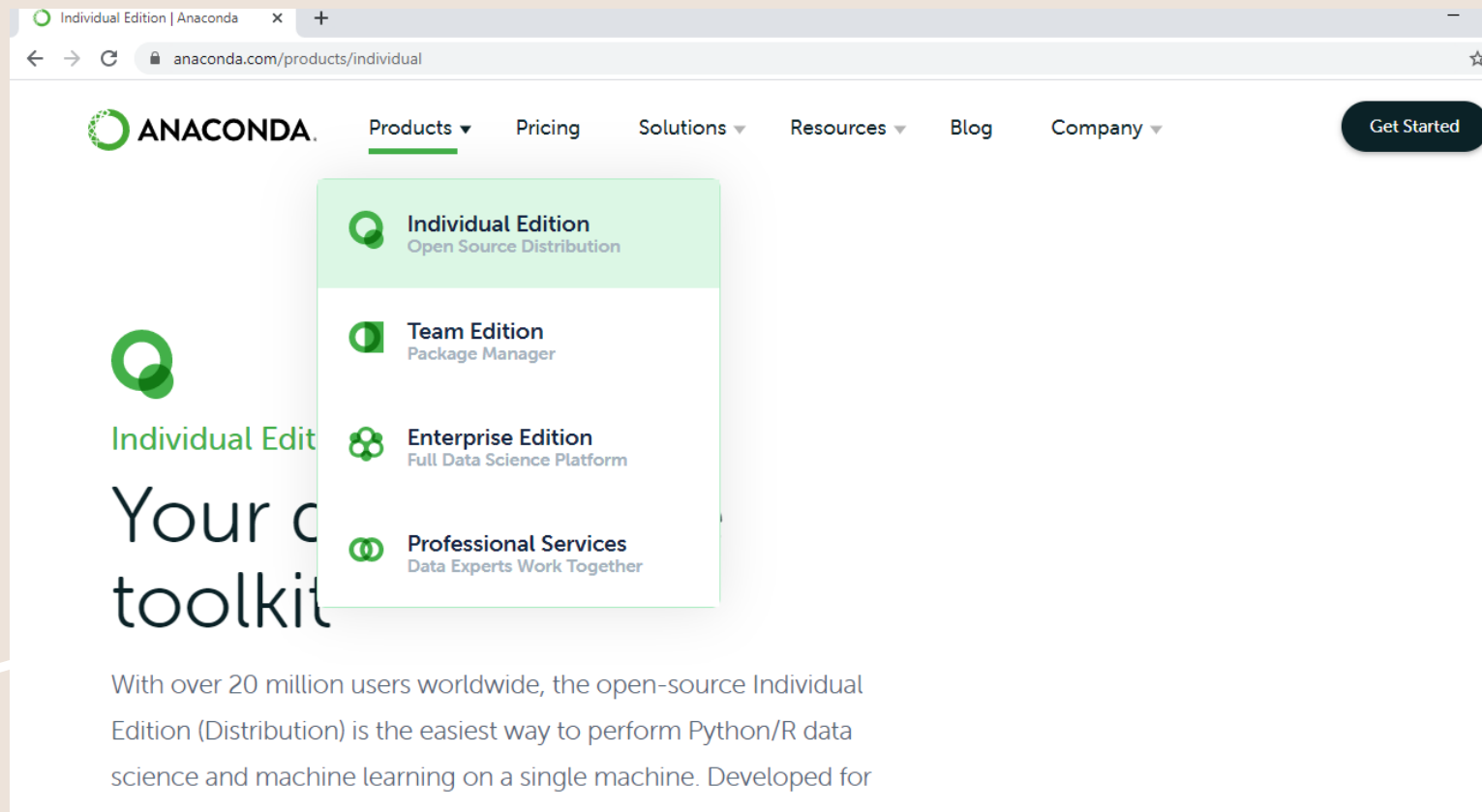
Guido Van Rossum

Softwares

- Basic python IDLE
 - from <https://www.python.org/downloads/>
- Jupyter Notebook by Anaconda Distributions
 - From <https://www.anaconda.com/products/individual>
- Google Colab by Google cloud service
 - From <https://colab.research.google.com/>
- Different online editors
 - From <https://repl.it/languages/python3>
 - Kaggle
 - Azure Jupyter notebooks

Anaconda Installation

Installation








The screenshot shows the Anaconda website's product page for Individual Edition. The browser's address bar displays 'anaconda.com/products/individual'. The navigation bar includes the Anaconda logo, a 'Products' dropdown menu (which is open), and links for 'Pricing', 'Solutions', 'Resources', 'Blog', and 'Company'. A 'Get Started' button is located in the top right corner. The 'Products' dropdown menu lists four options: 'Individual Edition' (Open Source Distribution), 'Team Edition' (Package Manager), 'Enterprise Edition' (Full Data Science Platform), and 'Professional Services' (Data Experts Work Together). The main content area features the Anaconda logo and the text 'Individual Edition Your data toolkit'. Below this, a paragraph states: 'With over 20 million users worldwide, the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for'.

Individual Edition | Anaconda x +

anaconda.com/products/individual ☆

ANACONDA Products ▾ Pricing Solutions ▾ Resources ▾ Blog Company ▾ [Get Started](#)

-  **Individual Edition**
Open Source Distribution
-  **Team Edition**
Package Manager
-  **Enterprise Edition**
Full Data Science Platform
-  **Professional Services**
Data Experts Work Together

 Individual Edition
Your data toolkit

With over 20 million users worldwide, the open-source Individual Edition (Distribution) is the easiest way to perform Python/R data science and machine learning on a single machine. Developed for

Downloading Anaconda Software

Windows

Python 3.7

64-Bit Graphical Installer (466 MB)

32-Bit Graphical Installer (423 MB)

Python 2.7

64-Bit Graphical Installer (413 MB)

32-Bit Graphical Installer (356 MB)

MacOS

Python 3.7

64-Bit Graphical Installer (442 MB)

64-Bit Command Line Installer (430 MB)

Python 2.7

64-Bit Graphical Installer (637 MB)

64-Bit Command Line Installer (409 MB)

Linux

Python 3.7

64-Bit (x86) Installer (522 MB)

64-Bit (Power8 and Power9) Installer (276 MB)

Python 2.7

64-Bit (x86) Installer (477 MB)

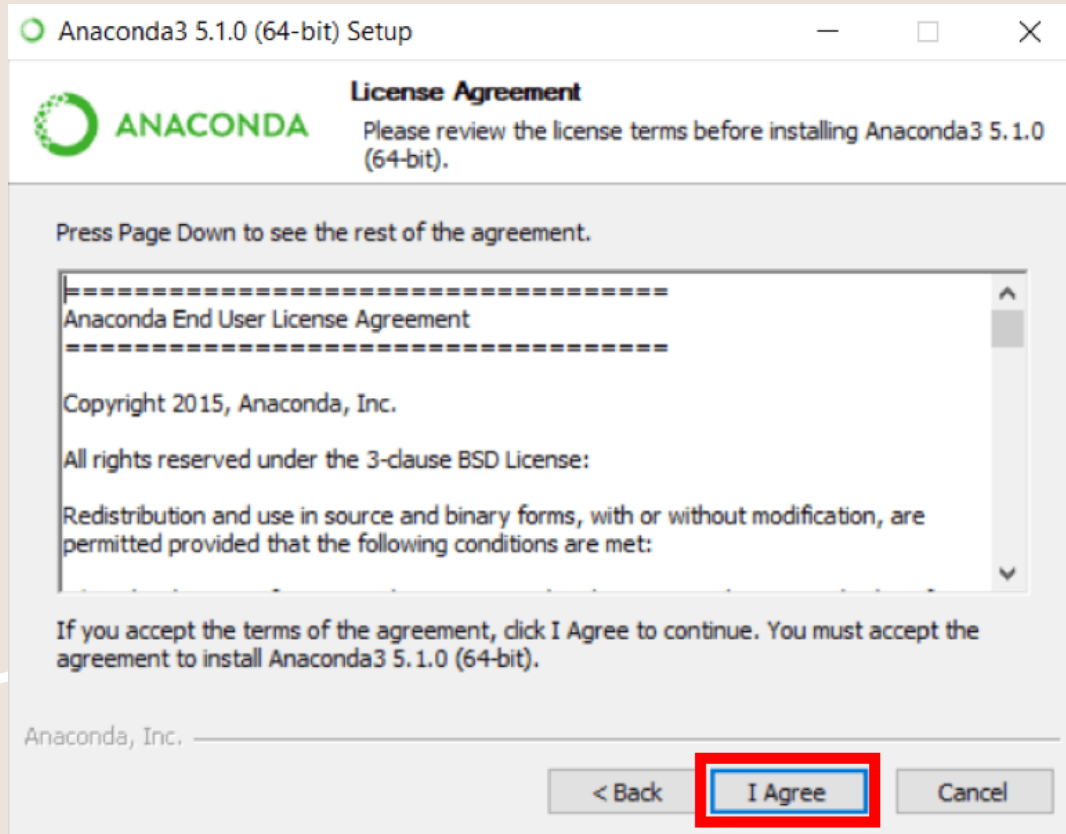
64-Bit (Power8 and Power9) Installer (295 MB)

Visit: <https://www.anaconda.com/products/individual>

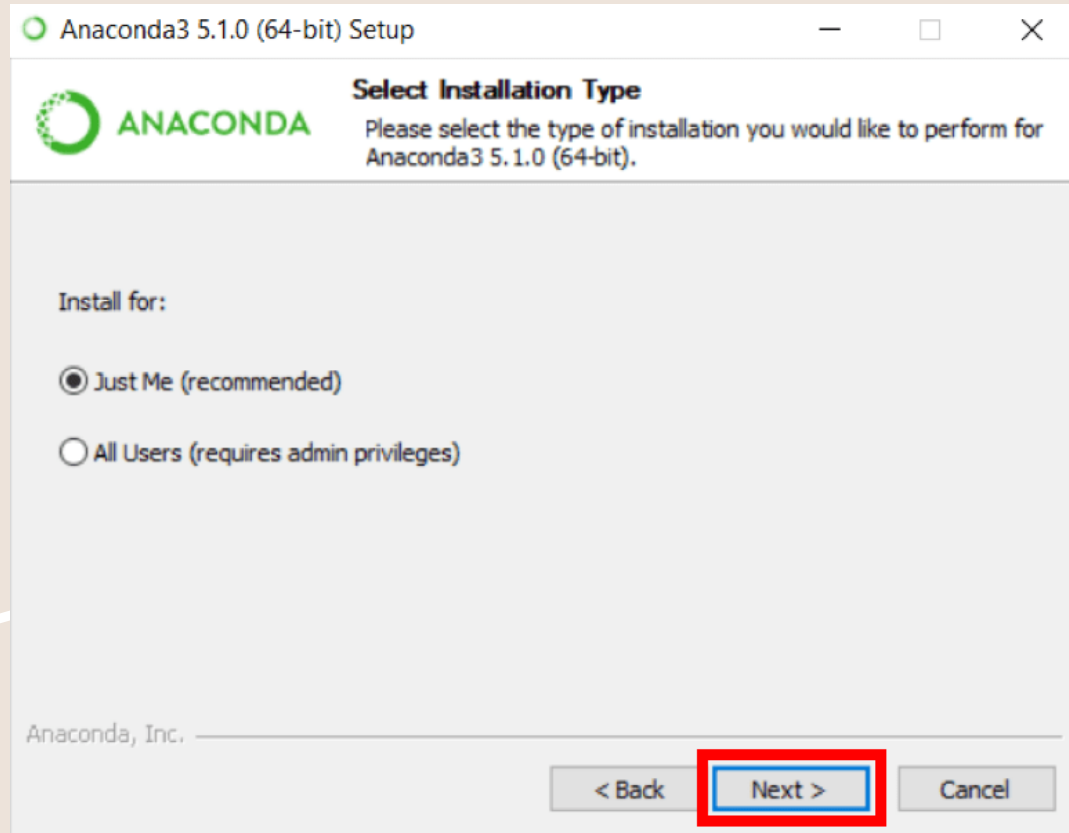
Installation



Installation




Installation



Installation

Anaconda3 5.1.0 (64-bit) Setup

 **ANACONDA** **Choose Install Location**
Choose the folder in which to install Anaconda3 5.1.0 (64-bit).

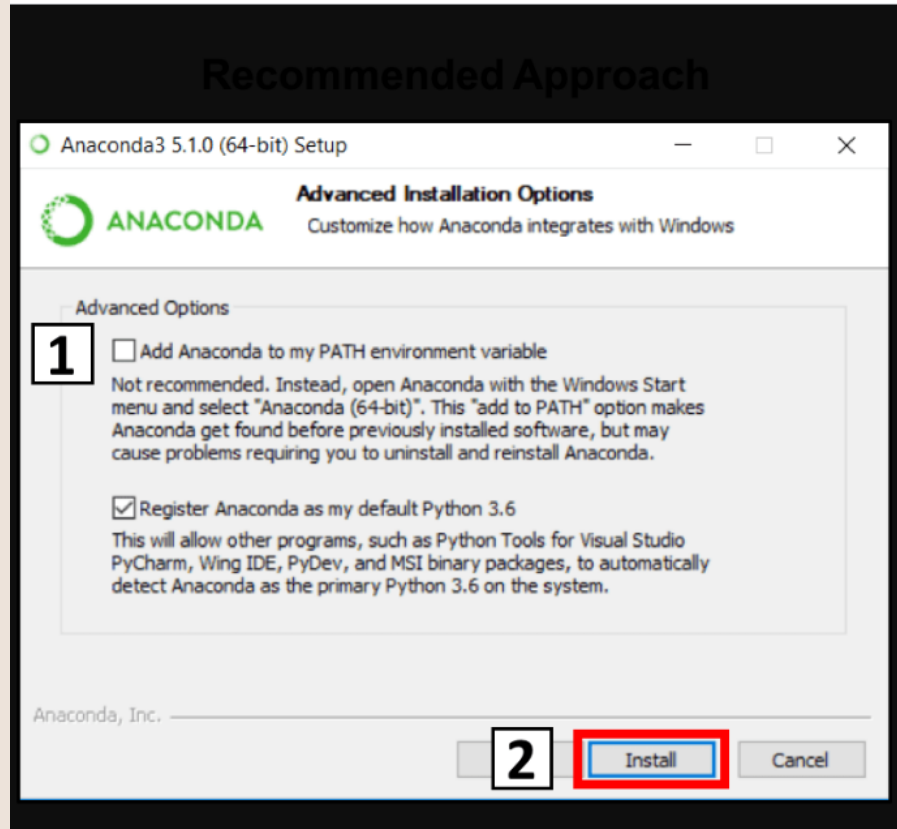
Setup will install Anaconda3 5.1.0 (64-bit) in the following folder. To install in a different folder, click Browse and select another folder. Click Next to continue.

Destination Folder

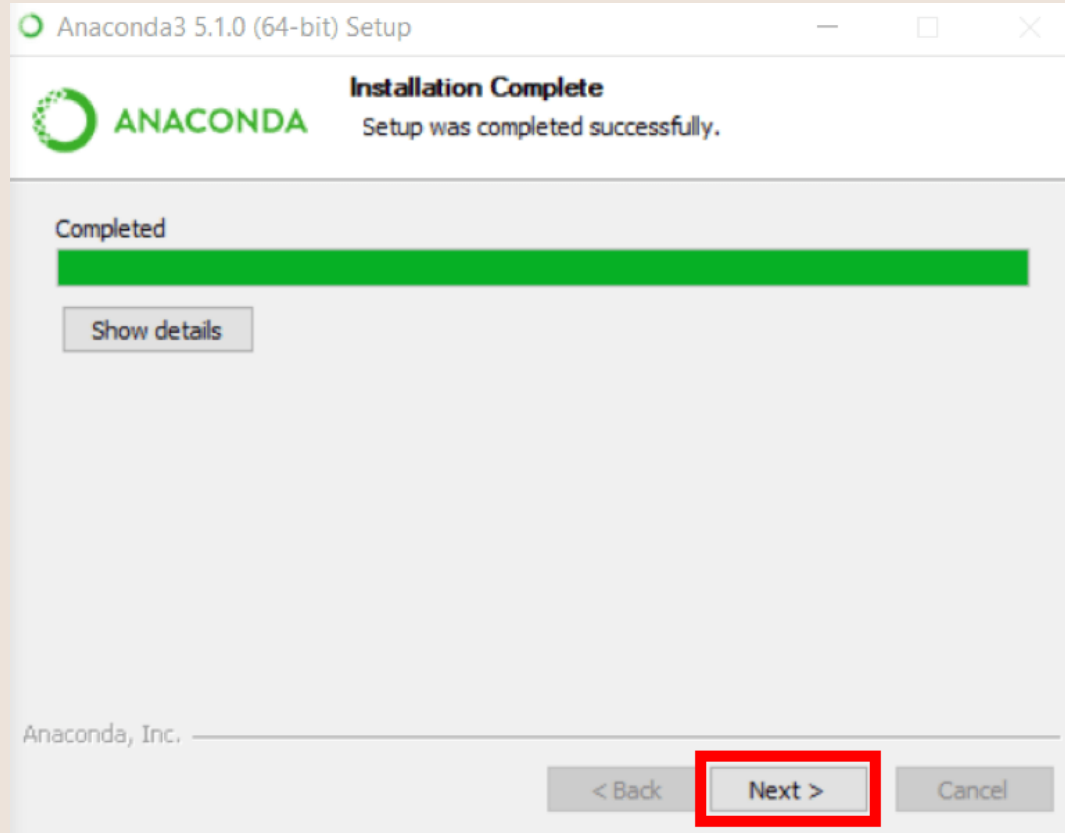
Space required: 2.5GB
Space available: 479.8GB

Anaconda, Inc.

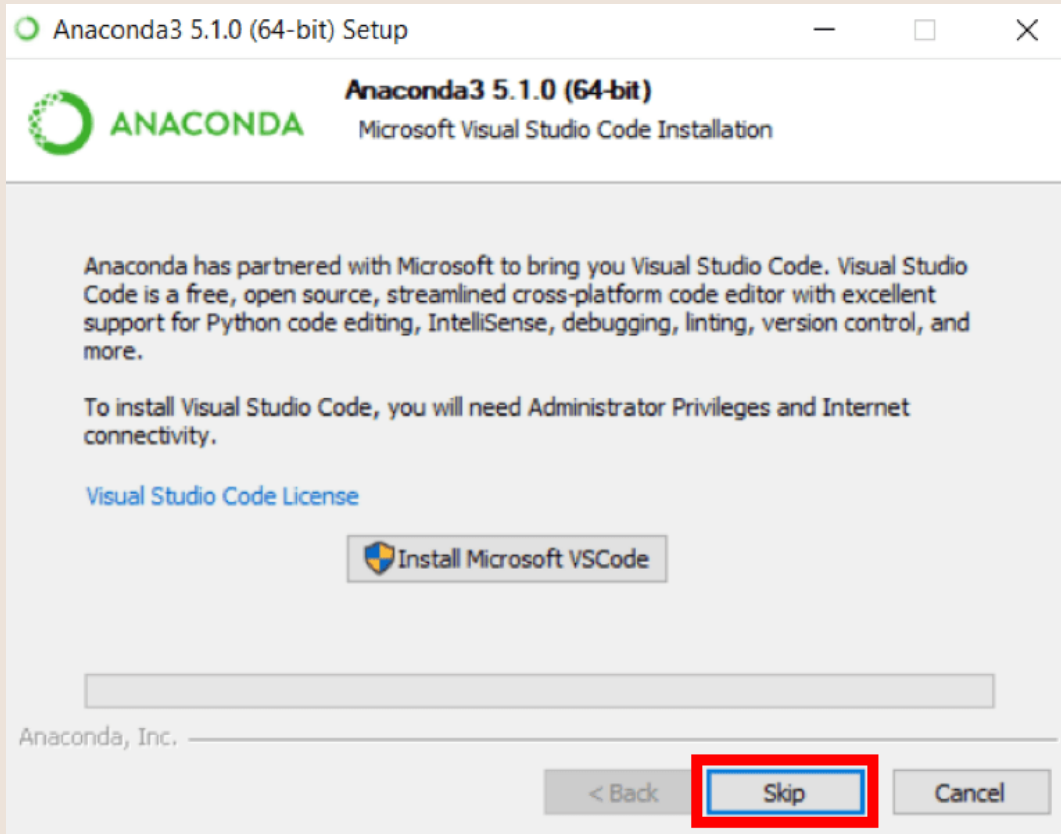
Installation



Installation

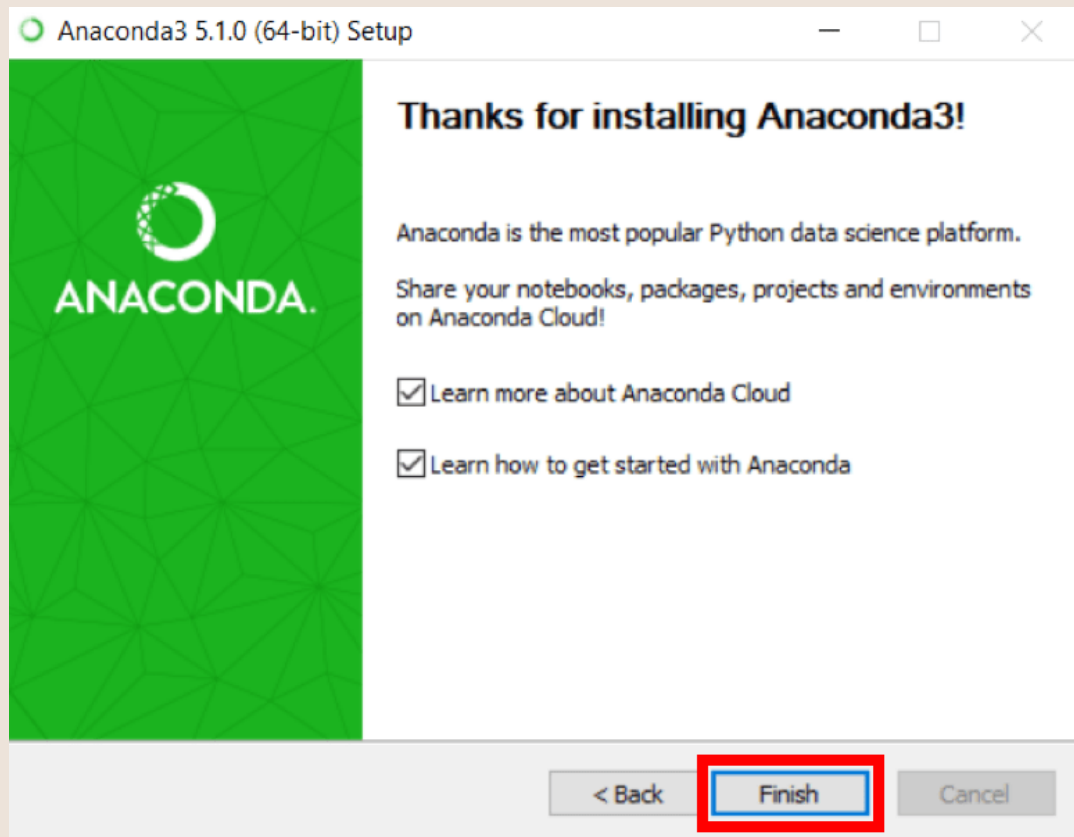


Installation



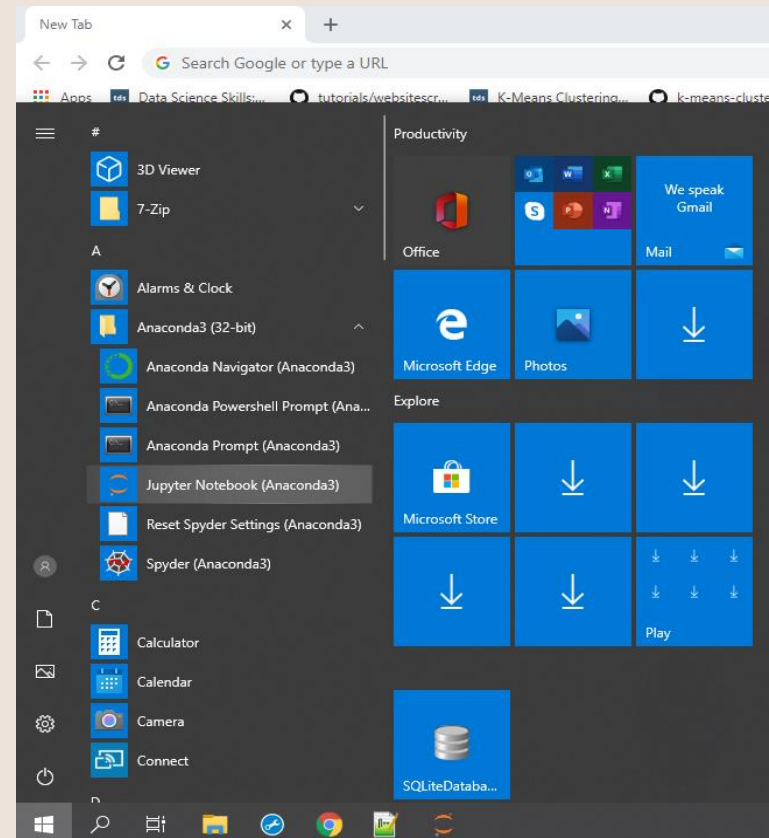
Instructions -> Login Name, Github
Why, Where
A

Installation



Let us start Jupyter Notebook

Launch Jupyter Notebook



It's time to experience  pythonTM