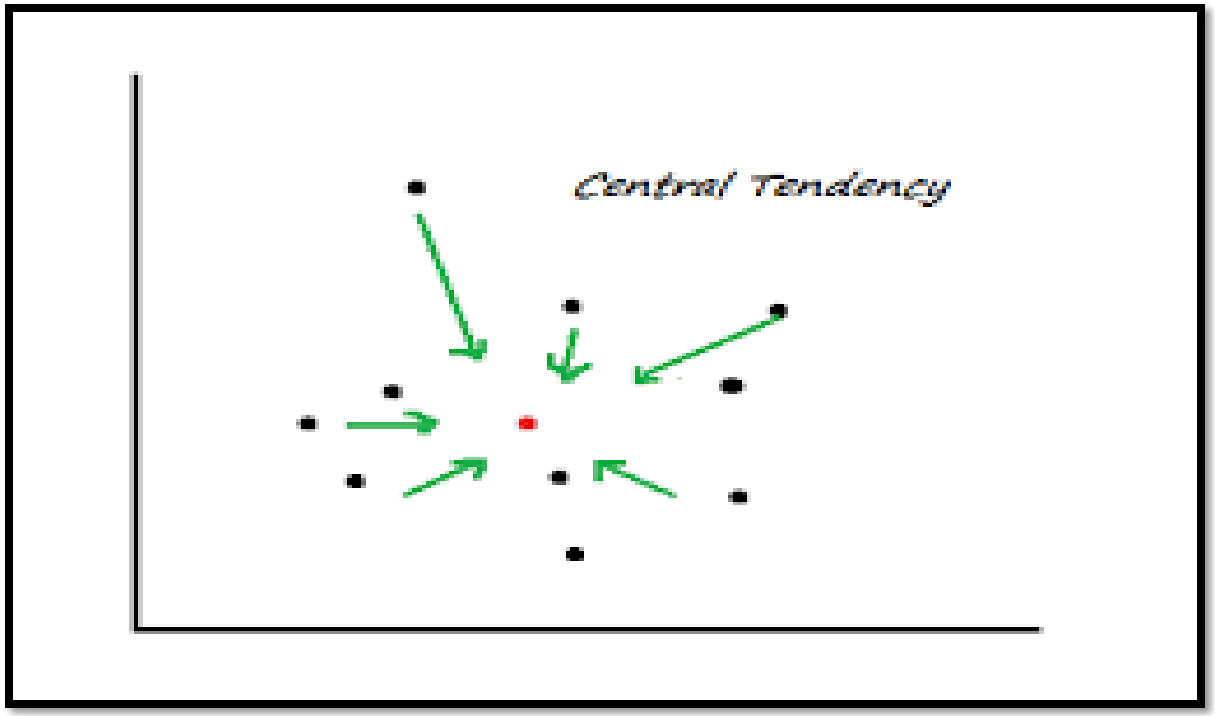
DMDW Lab using PYTHON

5th Semester
Department of Computer Science and Engineering
GIET University, Gunupur

Assignment 1

Measures of Central Tendency

- It describes distribution of data focusing on central location around which all other data are clustered.



Measures of Central Tendency

- It attempts to describe set of data by identifying the central position within which data is set.
- Measure of central tendency:
  1. Mean
  2. Median
  3. Mode

Mean

- The mean (or average) is the most popular and well known measure of central tendency. It can be used with both discrete and continuous data, although its use is most often with continuous data.
- The mean is equal to the sum of all the values in the data set divided by the number of values in the data set. So, if we have n values in a data set and they have values x1,x2, …,xn, the sample mean, usually denoted by $\bar{x}$ (pronounced "x bar"), is:

$$\bar{x} = \frac{x1,x2,\ \ldots,xn}{n}$$

- This formula is usually written in a slightly different manner using the Greek capitol letter, ∑, pronounced "sigma", which means "sum of...":

$$\bar{x} = \frac{\sum x}{n}$$

Example: For example, consider the wages of staff at a factory below

Staff	1	2	3	4	5	6	7	8	9	10
Salary	15k	18k	16k	14k	15k	15k	12k	17k	90k	95k

The mean salary for these ten staff is $30.7k.

Median

- The median is the middle score for a set of data that has been arranged in order of magnitude.
- The median is less affected by outliers and skewed data. In order to calculate the median, suppose we have the data below
- Ex-1) 65 55 89 56 35 14 56 55 87 45 92

We first need to rearrange that data into order of magnitude

14 35 45 55 55 **56** 56 65 87 89 92

Our median mark is the middle mark - in this case is 56

- Ex-2) 65 55 89 56 35 14 56 55 87 45

We again rearrange that data into order of magnitude (smallest first):

14 35 45 55 **55** **56** 56 65 87 89

Only now we have to take the 5th and 6th score in our data set and average them to get a median of 55.5.

Mode

- The mode is the most frequent score in our data set. On a histogram it represents the highest bar in a bar chart or histogram in fig-1.

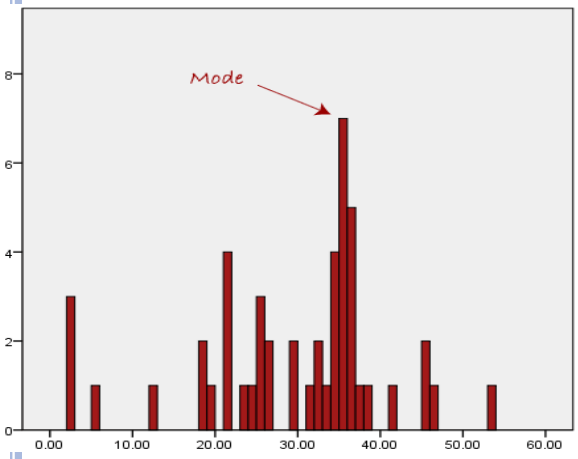


Fig-1

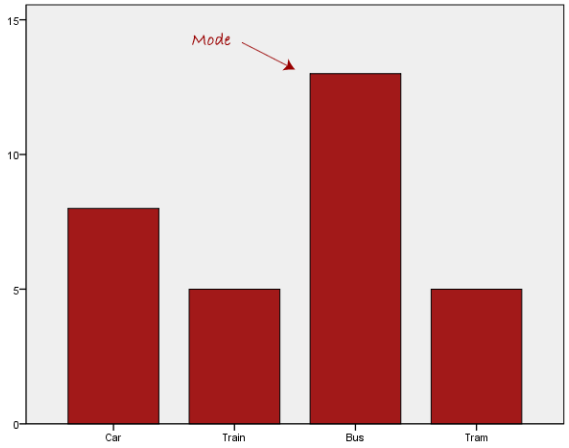


Fig-2

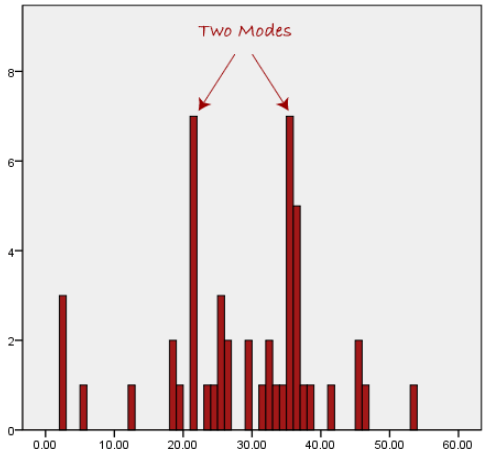
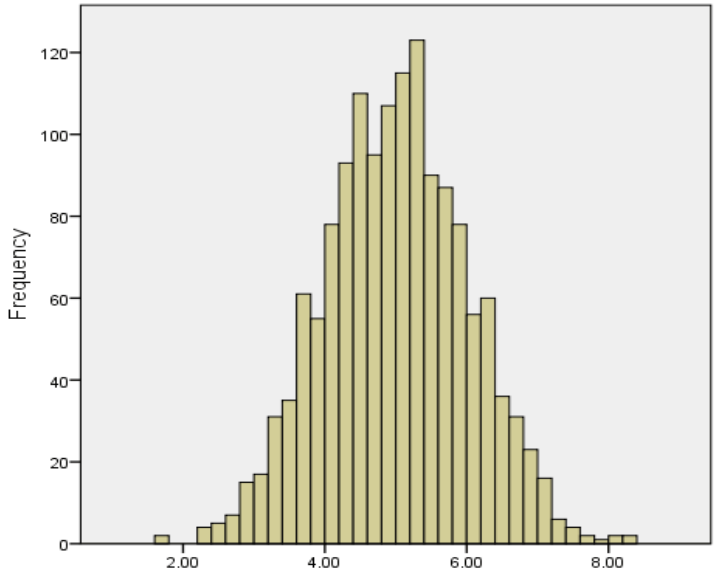


Fig-3

Normally, the mode is used for categorical data where we wish to know which is the most common category, as illustrated in fig-2.
However, one of the problems with the mode is that it is not unique, so it leaves us with problems when we have two or more values that share the highest frequency, such as fig-3.

Skewed Distributions

- An example of a normally distributed set of data is presented below.

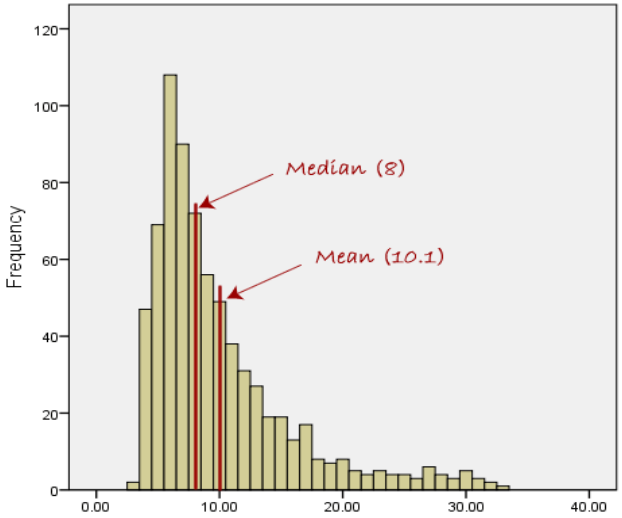


•In any symmetrical distribution the mean, median and mode are equal.

• **Mean** is widely preferred as the best measure of central tendency because it is the measure that includes all the values in the data set for its calculation.

CONTD.

However, when our data is skewed, for example, as with the right-skewed data set below:



•Median is generally considered to be the best representative of the central location of the data.
•The more skewed the distribution, the greater the difference between the median and mean .
•The greater emphasis should be placed on using the median as opposed to the mean.

Summary of when to use the mean, median and mode

Please use the following summary table to know what the best measure of central tendency is with respect to the different types of variable.

Type of Variable	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Interval/Ratio (not skewed)	Mean
Interval/Ratio (skewed)	Median

Variance and Standard Deviation

$$Variance, \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$Standard\ Deviation, \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}}$$

Where $x_i = data\ set\ values$

$\bar{x} = mean\ of\ the\ data\ set$

Example

The ages of you and your friends are 25, 26, 27, 30, and 32.

First, we must find the mean age: (25 + 26 + 27 + 30 + 32) / 5 = 28.

Then, we need to calculate the differences from the mean for each of the 5 friends.

25 – 28 = -3
26 – 28 = -2
27 – 28 = -1
30 – 28 = 2
32 – 28 = 4

Next, to calculate the variance, we take each difference from the mean, square it, then average the result.

Variance = $((-3)^2 + (-2)^2 + (-1)^2 + 2^2 + 4^2)/ 5$

$= (9 + 4 + 1 + 4 + 16) / 5 = 6.8$

Variance is 6.8. Standard deviation is the square root of the variance, which is 2.61.

PRACTICE-1

- Write the python code for following statistical operations with and without library function:
  - ✓ Mean
  - ✓ Median
  - ✓ Mode
  - ✓ Standard Deviation and
  - ✓ Variance

Mean Without Library Function

- # Mean without using library

```
n_num = [1, 2, 3, 4, 5]
n = len(n_num)
get_sum = sum(n_num)
mean = get_sum / n
print("Mean / Average is: " + str(mean))
```

Median without library function

```
# Median without using library
n_num = [1, 2, 3, 4, 5]
n = len(n_num)
n_num.sort()
 if n % 2 == 0:
median1 = n_num[n//2]
median2 = n_num[n//2 - 1]
median = (median1 + median2)/2
else:
median = n_num[n//2]
print("Median is: " + str(median))
```

Mode without library function

```
# Python program to print mode of elements
from collections import Counter
  n_num = [1, 2, 3, 4, 5, 5]
n = len(n_num)
 data = Counter(n_num)
get_mode = dict(data)
mode = [k for k, v in get_mode.items() if v== max(list(data.values()))]
  if len(mode) == n:
    get_mode = "No mode found"
else:
      get_mode = "Mode is / are: " + ', '.join(map(str, mode))
 print(get_mode)
```

Mode with library function

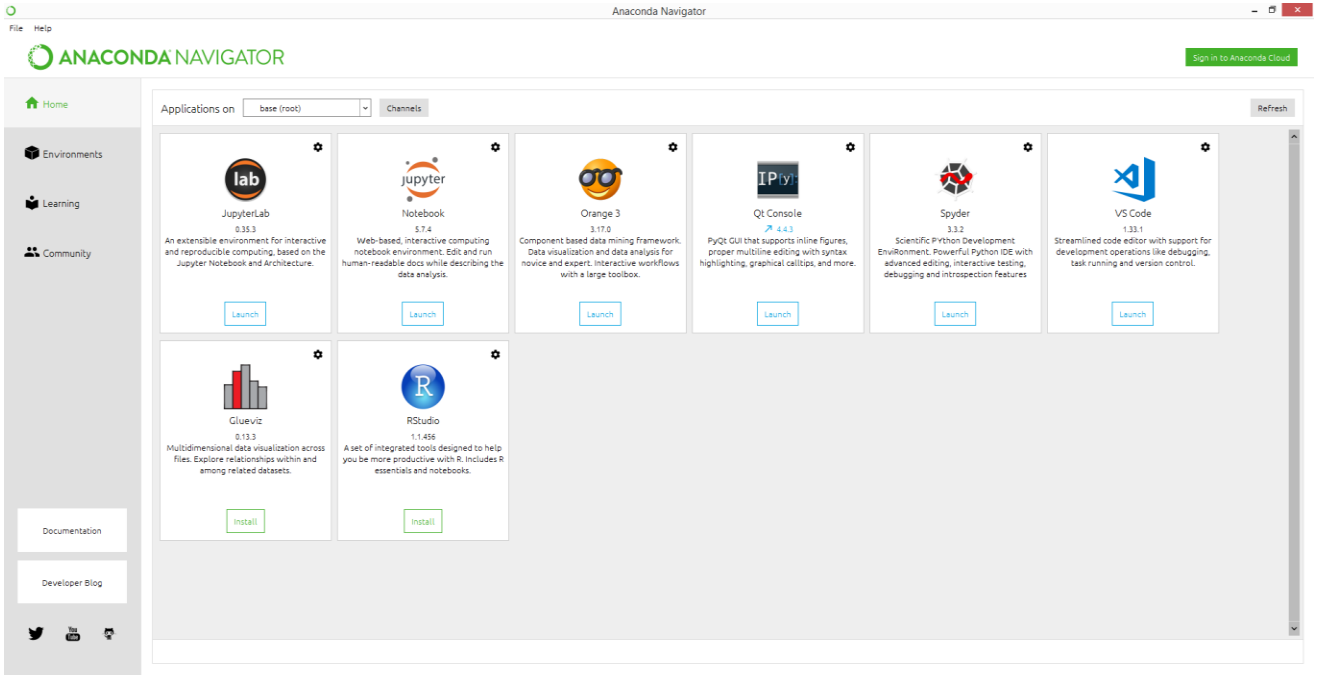
```
import numpy
speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]
x = numpy.mean(speed)
y = numpy.median(speed)
s = numpy.std(speed)
v = numpy.var(speed)
print(x)
print(y)
print(s)
print(v)
```

Mode with library function

```
from scipy import stats
speed = [99,86,87,88,111,86,103,87,94,78,77,85,86]
x = stats.mode(speed)
print(x)
```

Anaconda Platform

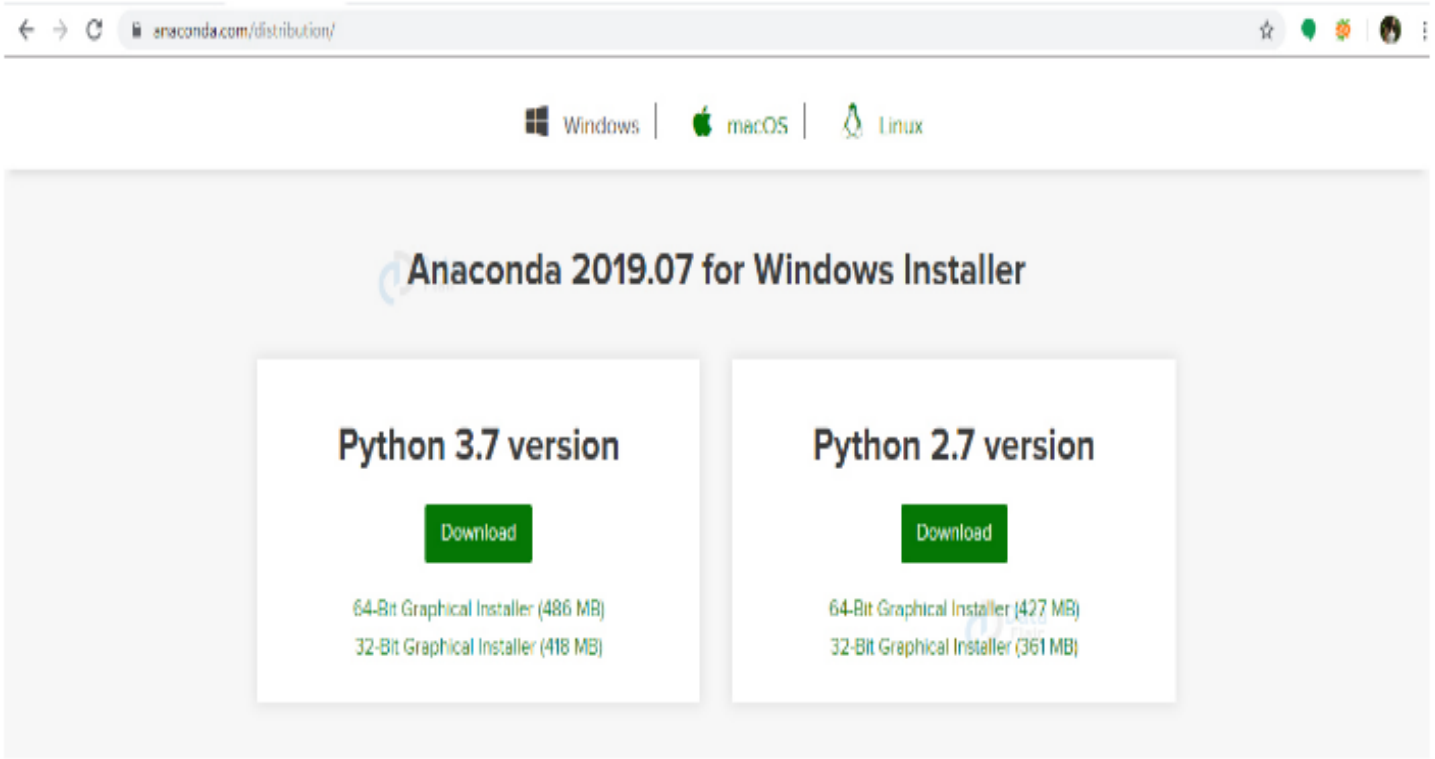
Anaconda Individual Edition is the world's most popular **Python** distribution platform with over 20 million users worldwide.



- Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands.

Anaconda Installation Steps

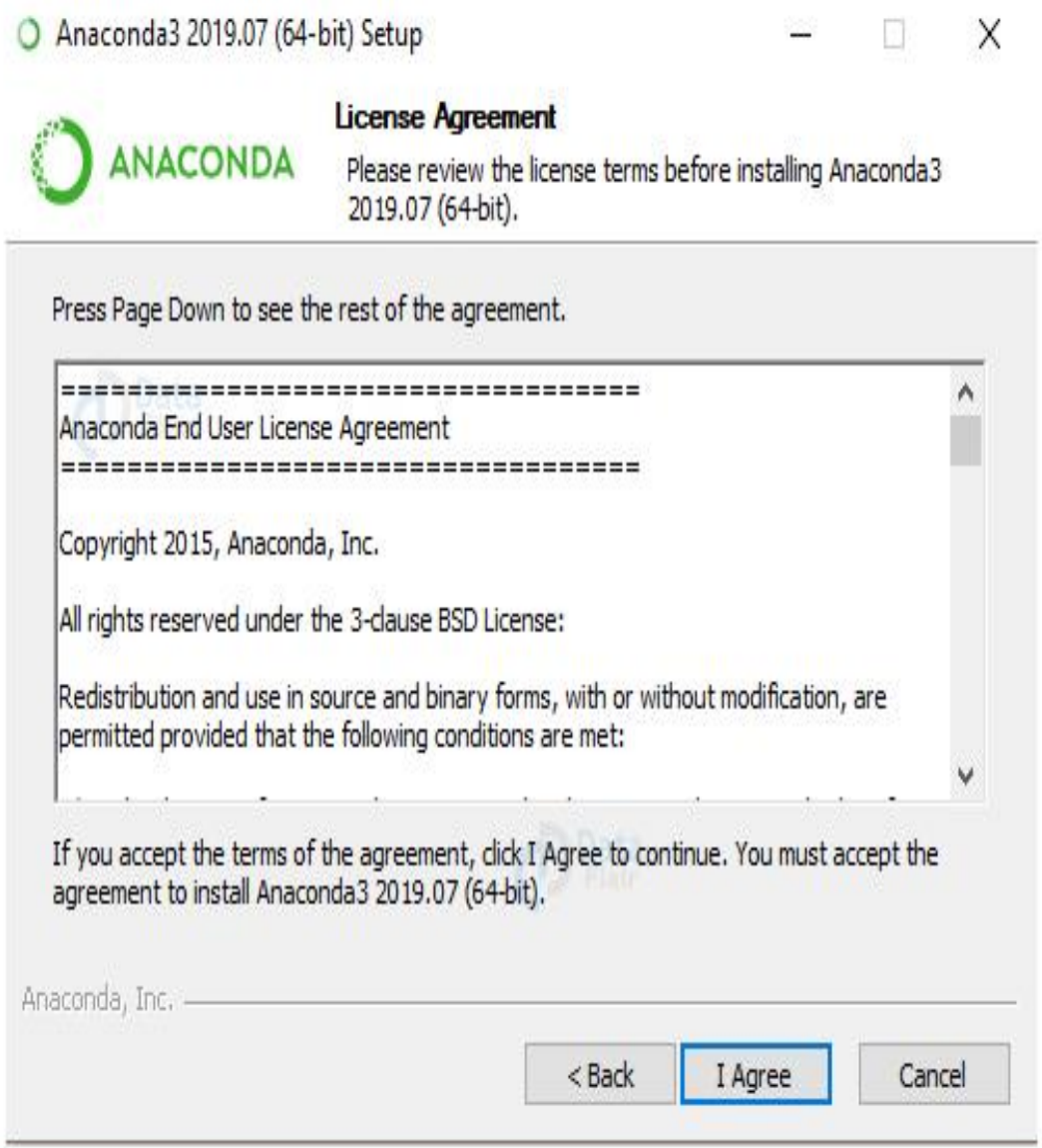
1. Go to this link and download Anaconda for Windows, Mac, or Linux: – Download anaconda



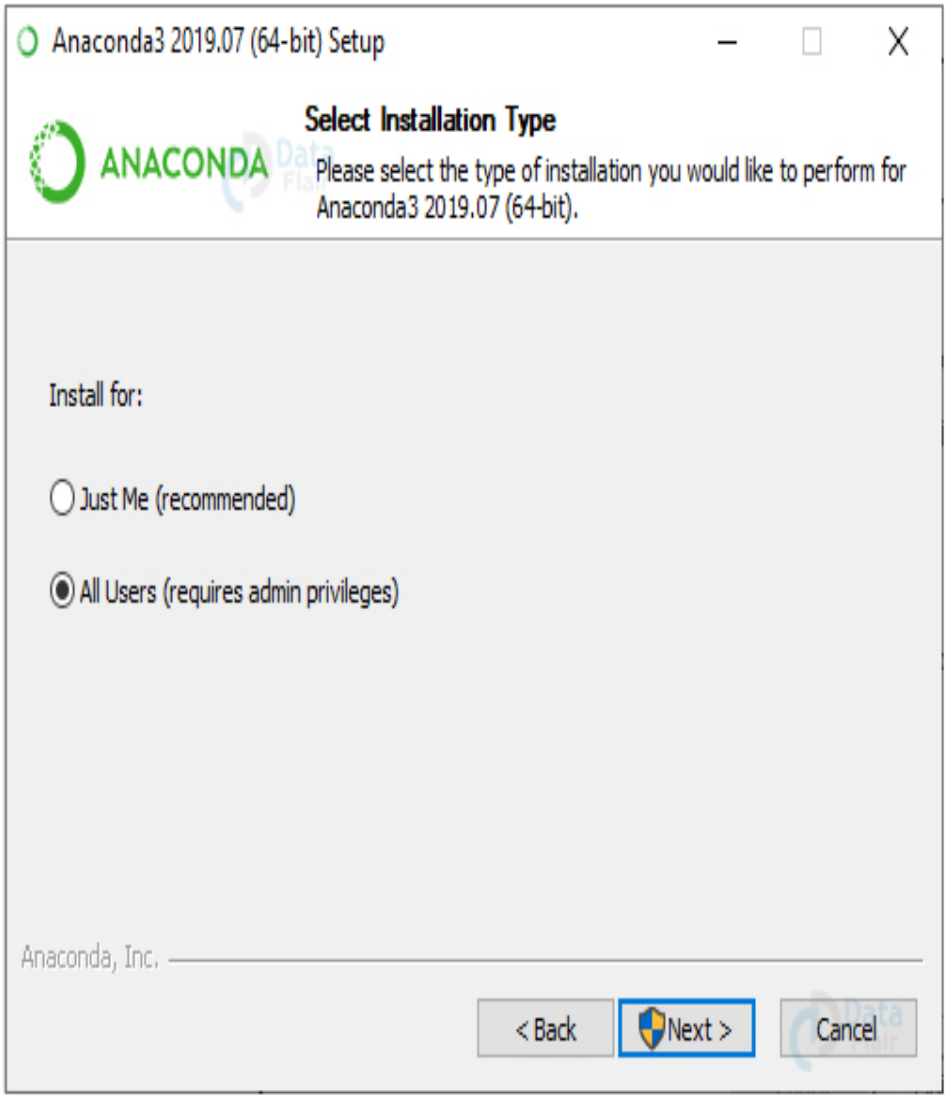
You can download the installer for Python 3.7 or for Python 2.7 (at the time of writing). And you can download it for a 32-bit or 64-bit machine.

2. Click on the downloaded .exe to open it. This is the Anaconda setup. Click next.

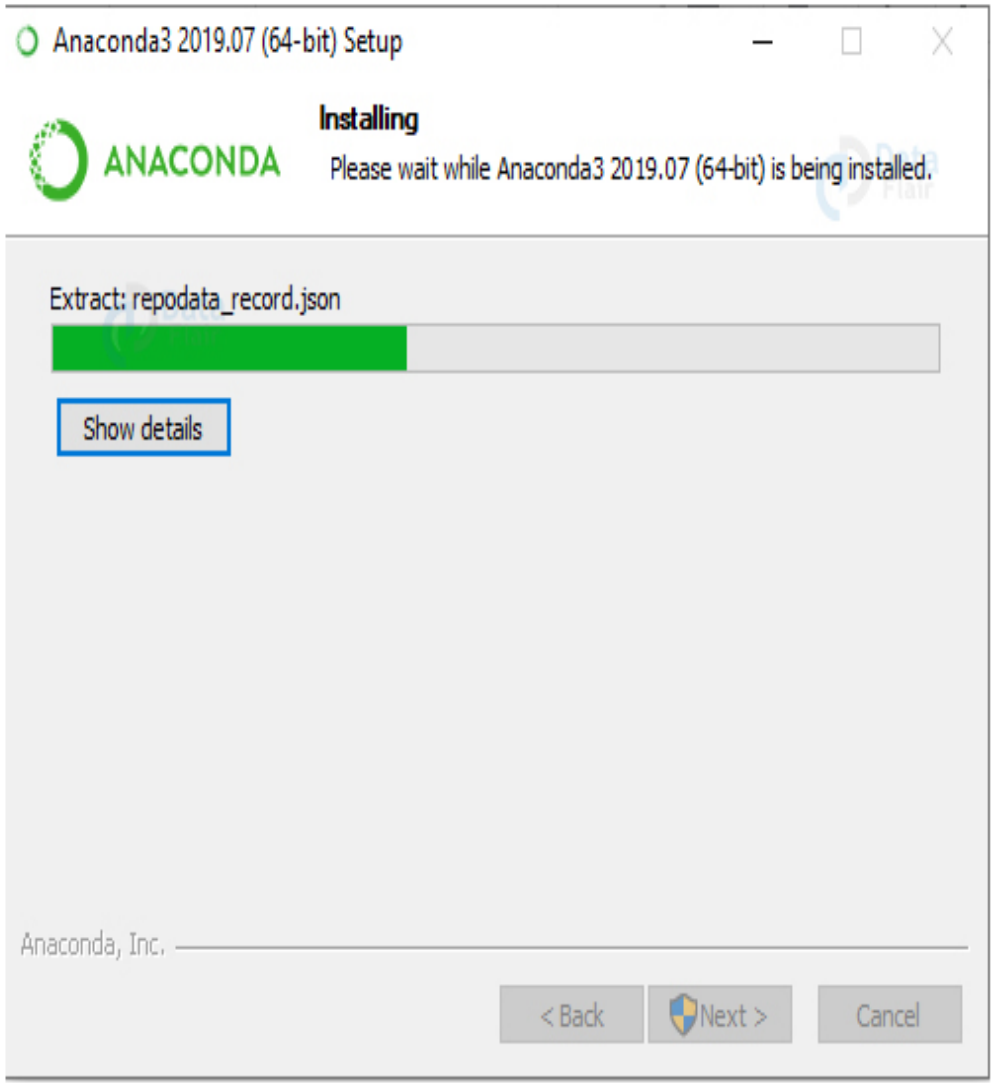
3. Now, you'll see the license agreement. Click on 'I Agree'.



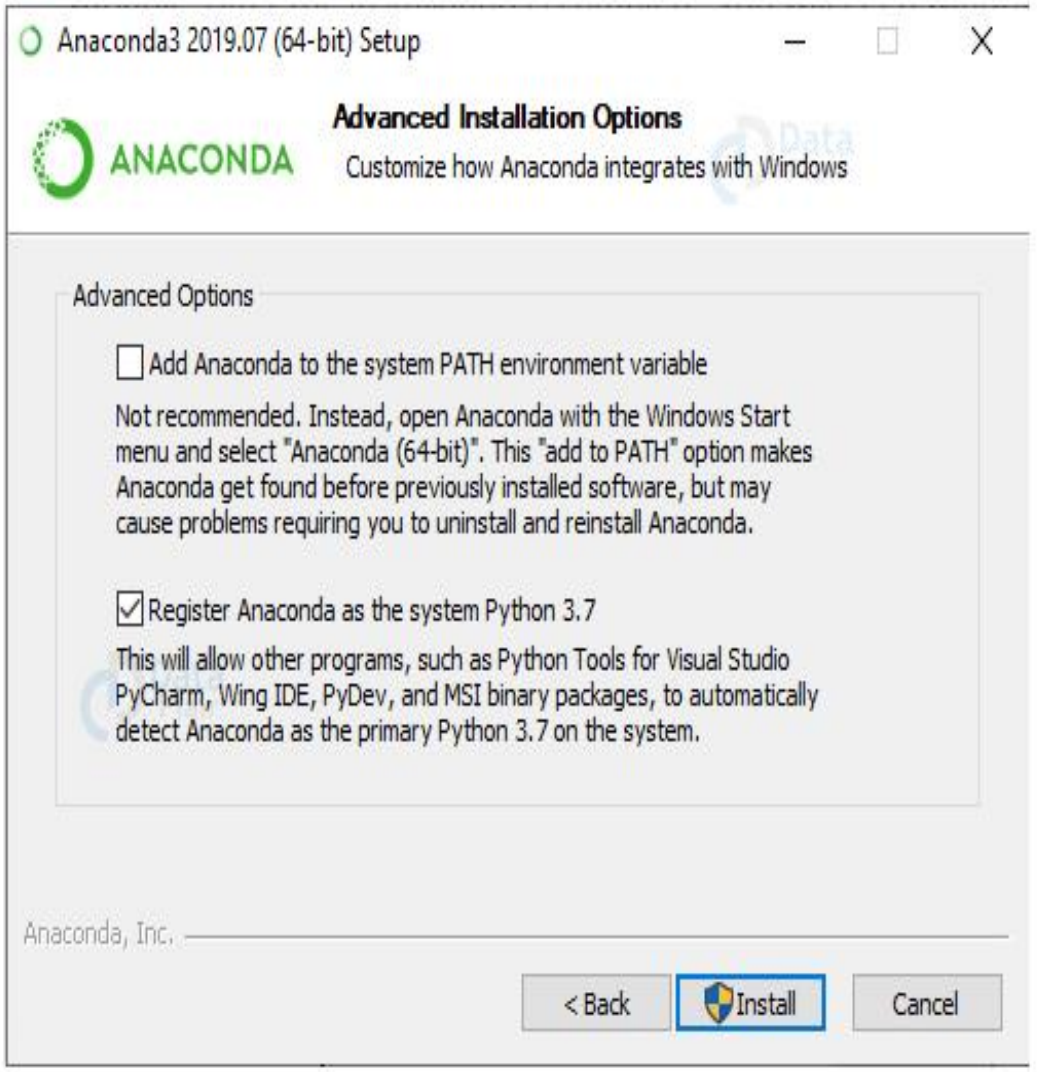
4. You can install it for all users or just for yourself. If you want to install it for all users, you need administrator privileges.



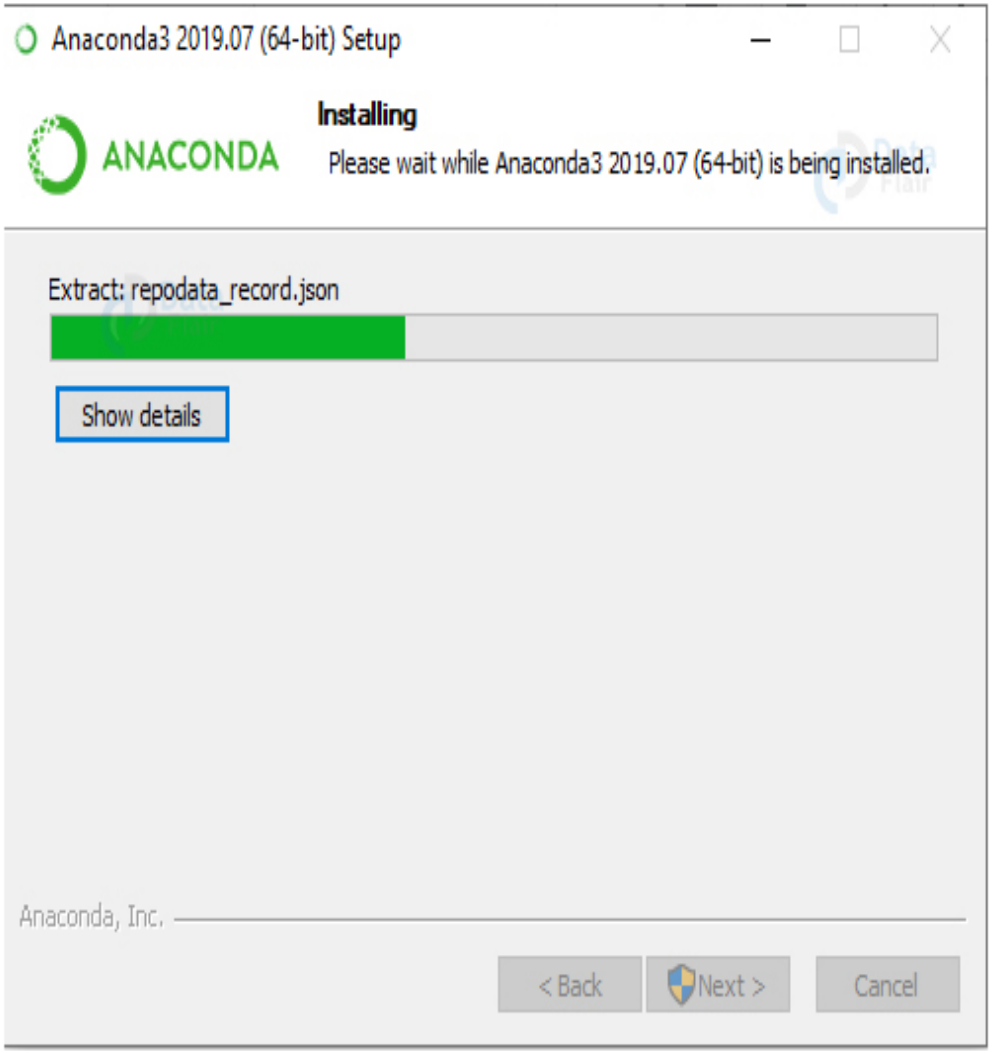
5. Choose where you want to install it. Here, you can see the available space and how much you need.



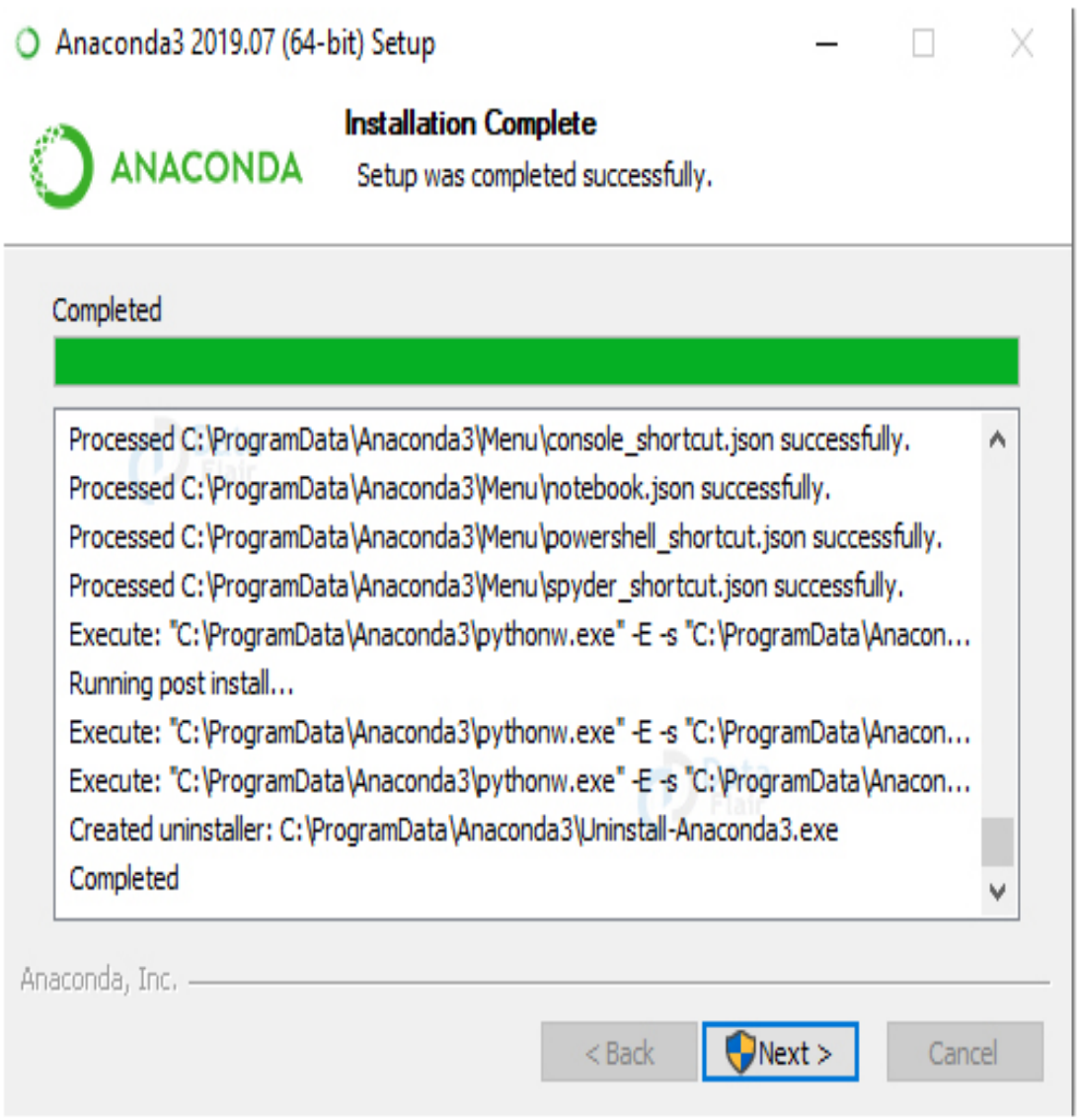
6. Now, you'll get some advanced options. You can add Anaconda to your system's PATH environment variable, and register it as the primary system Python 3.7. If you add it to PATH, it will be found before any other installation. Click on 'Install'.



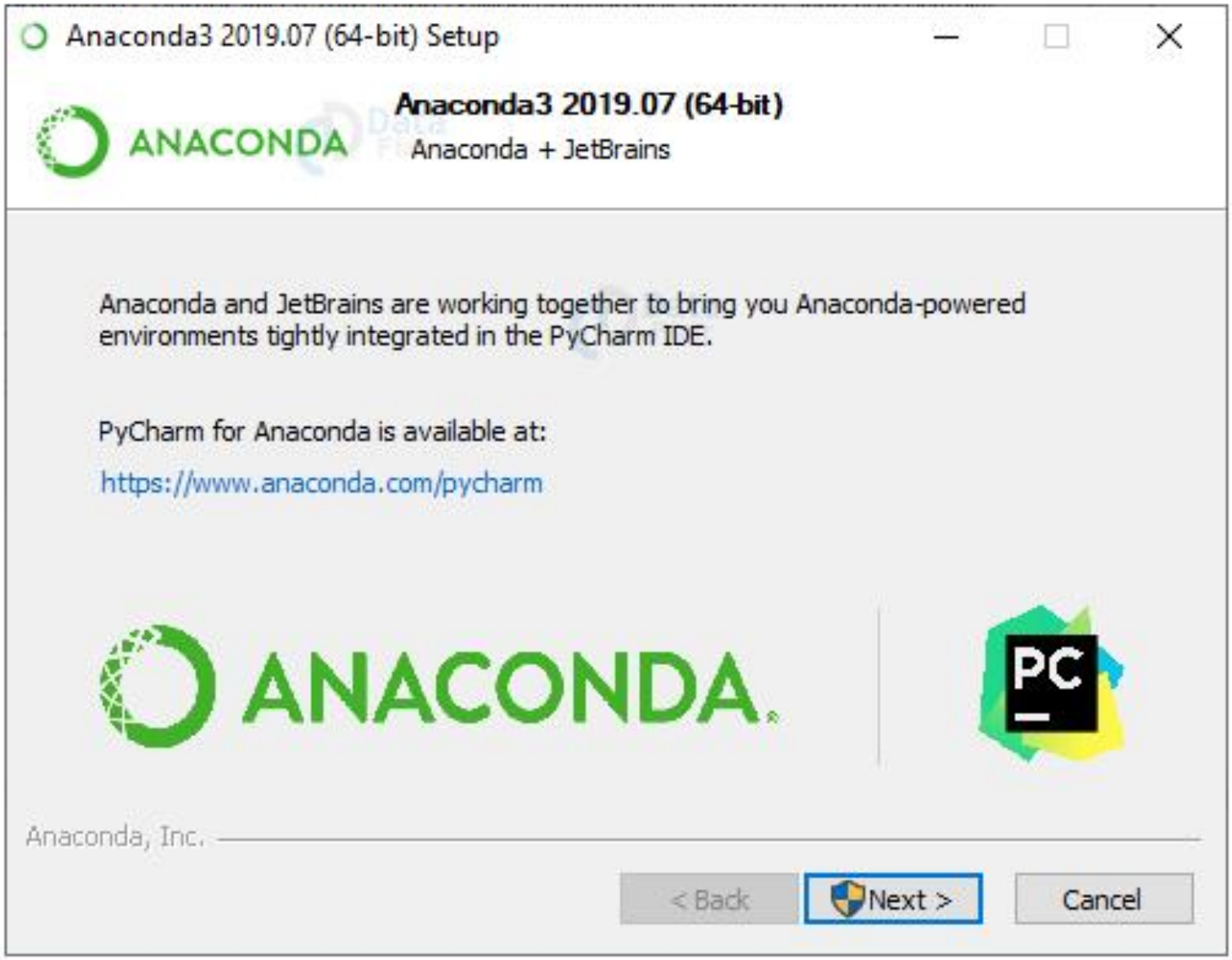
7. It will unpack some packages and extract some files on your machine. This will take a few minutes.



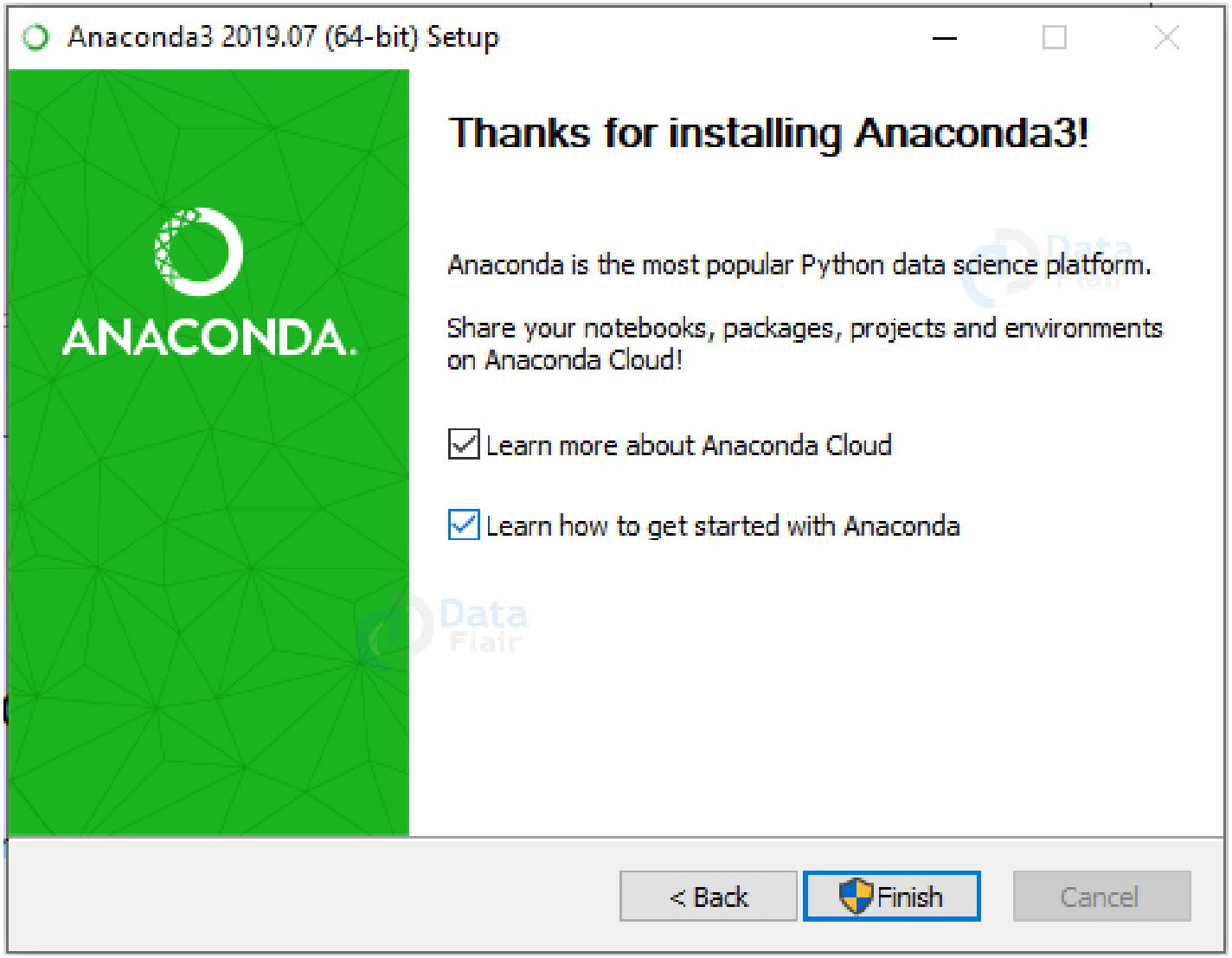
8. The installation is complete. Click Next.



9. This screen will inform you about PyCharm. Click Next.



10. The installation is complete. You can choose to get more information about Anaconda cloud and how to get started with Anaconda. Click Finish.

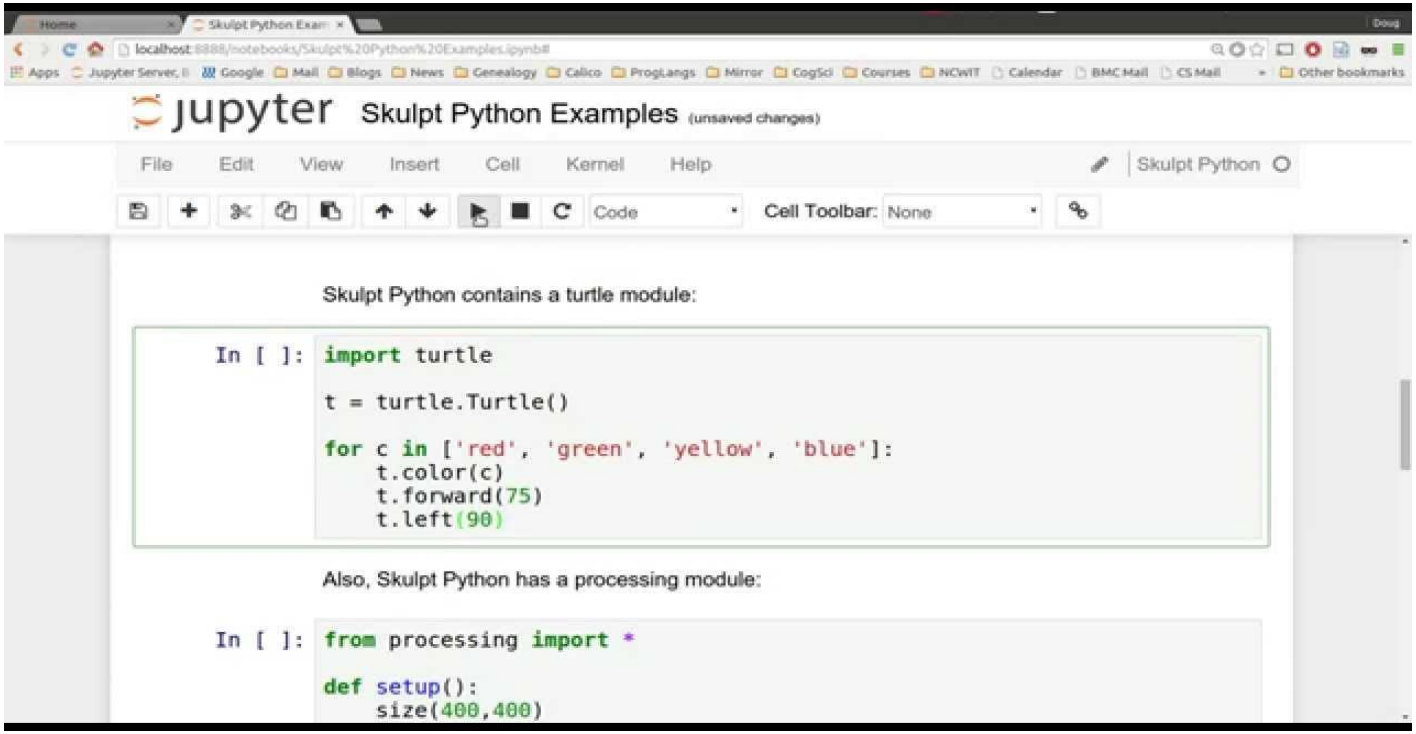


Benefits of Using Python Anaconda

- It is free and open-source
- It has more than 1500 Python/***R data science packages***
- It creates an environment that is easily manageable for deploying any project
- Download more than 1500 Python/R data science packages
- Manage libraries, dependencies, and environments with conda
- Build and train ML and deep learning models with scikit-learn, TensorFlow and Theano
- Use Dask, NumPy, Pandas and Numba to analyze data scalably and fast
- Perform visualization with Matplotlib, Bokeh, Datashader, and Holoviews

The Jupyter Notebook

- The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text.
- Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.



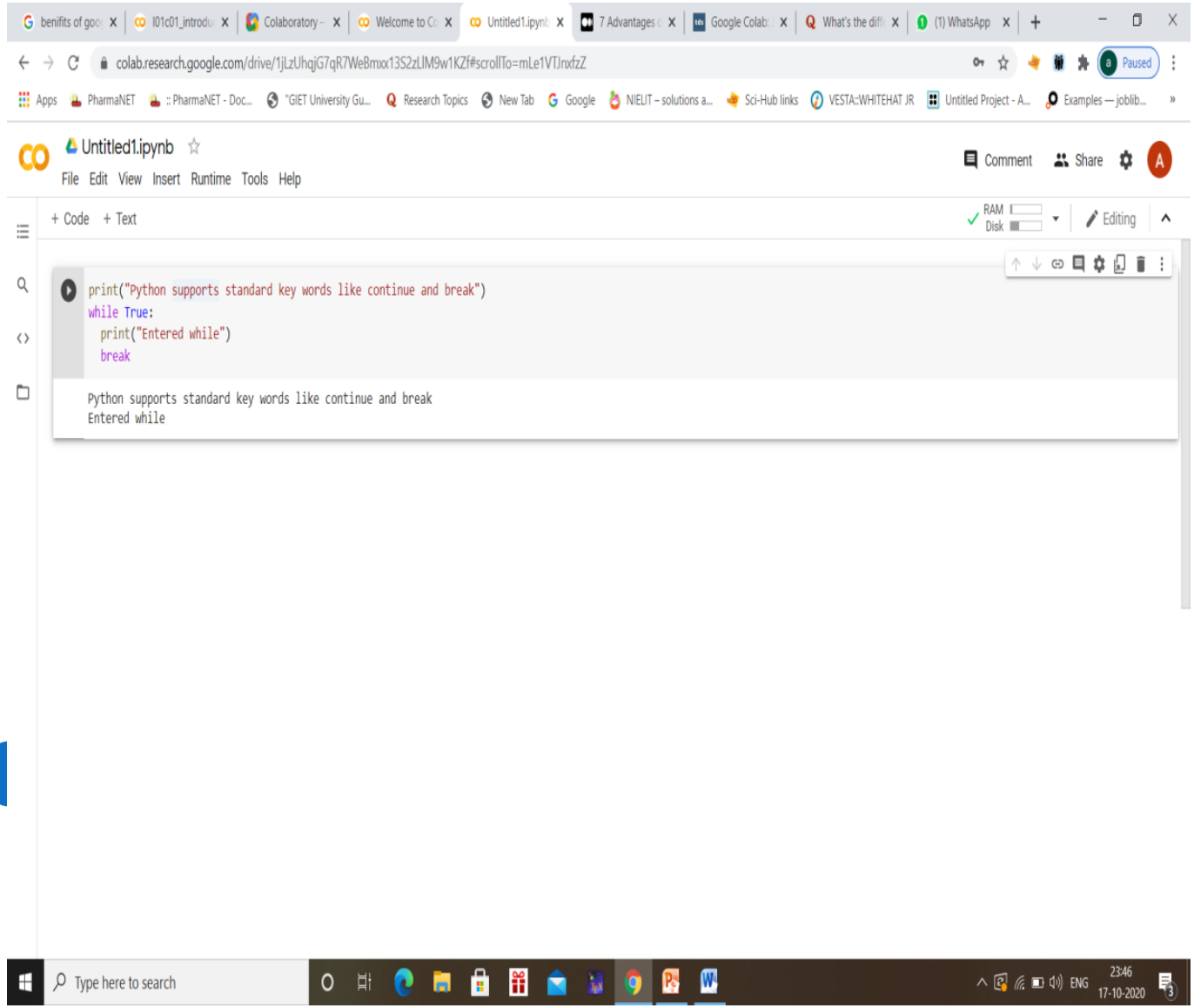
Introduction to Google-colab

Colaboratory, or '**Colab**' for short, allows you to write and execute Python in your browser, with

- ✓ Zero configuration required
- ✓ Free access to GPUs
- ✓ Easy sharing

Advantages

- It performs **all the tasks and code that Jupyter Notebook executes,** using Python 2 and 3.
- It is **THE Google Documents of Code**. The notebook can be shared and edited in real-time by different team members, add comments, see the edition history and go back to previous versions, like in google docs.
- **No more Anaconda.** It is all cloud-based and it doesn't require any main settings or installations. If the library that you want to use is not on Colab, just pip it as usual. Being installed in the virtual environment.
- **Personalization.** Add your own shortcuts, night/light/adaptive - mode, and fonts.
- **Playground mode.** With 2 clicks you can enter open a new notebook that won't be saved, and try different code options without affecting your original code.


colab.research.google.com/drive/1jLzUhqjG7qR7WeBmxx13S2zLIM9w1KZf#scrollTo=mLe1VTJnxfzZ
Untitled1.ipynb
File Edit View Insert Runtime Tools Help
Comment
Share
+ Code + Text
RAM
Disk
Editing
print("Python supports standard key words like continue and break")
while True:
print("Entered while")
break
Python supports standard key words like continue and break
Entered while
Type here to search
ENG
23:46
17-10-2020

Assignment Question

1. Write a python code for finding mean, median and mode .
2. Write a python code for calculating variance and standard deviation for the set of elements.
3. Practice some basic python programs.