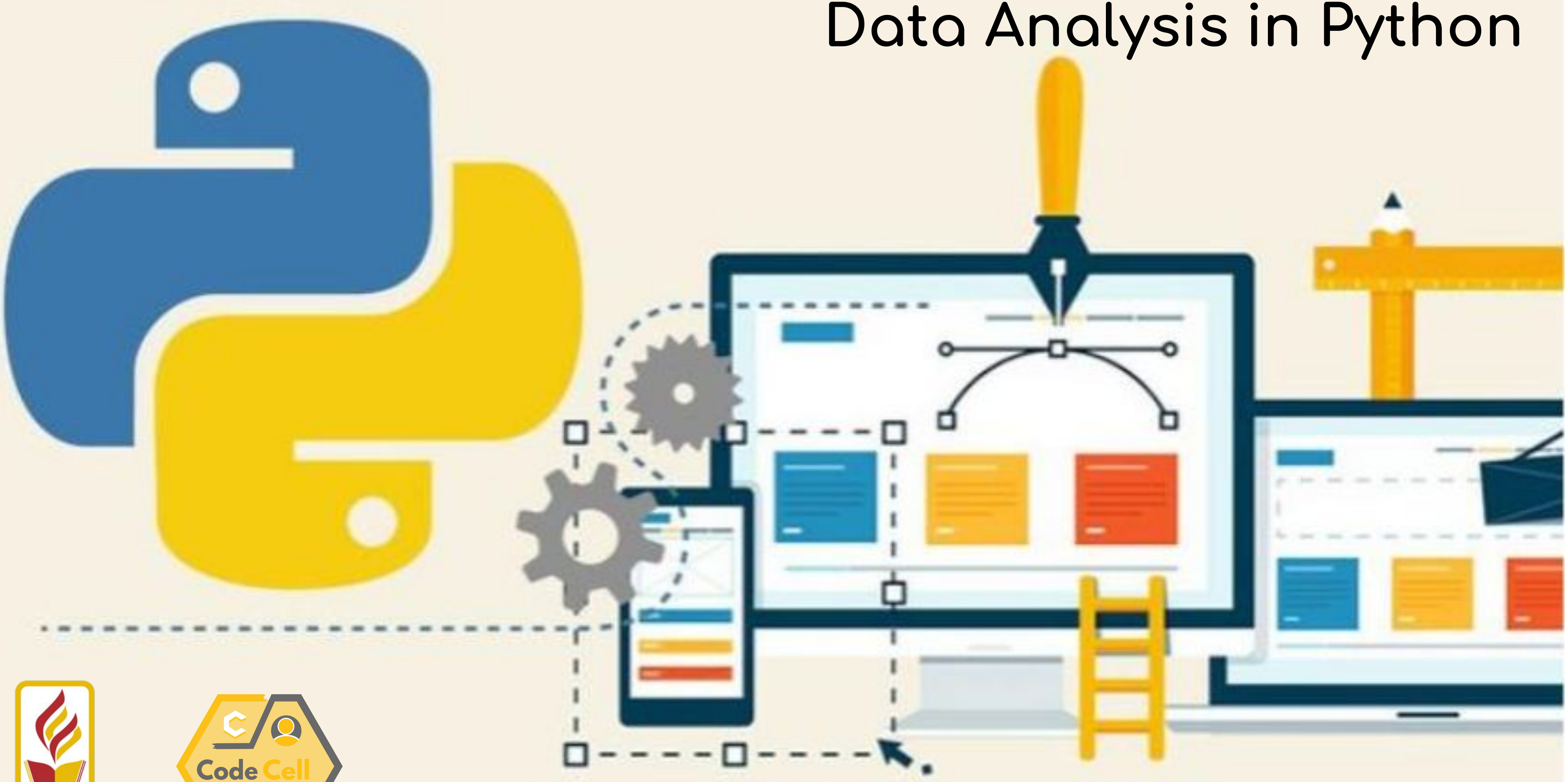
Data Analysis in Python



V.E.S.
Since 1962

Code Cell
Tinker's from
Computer Department
VESIT

What is Data Analysis?

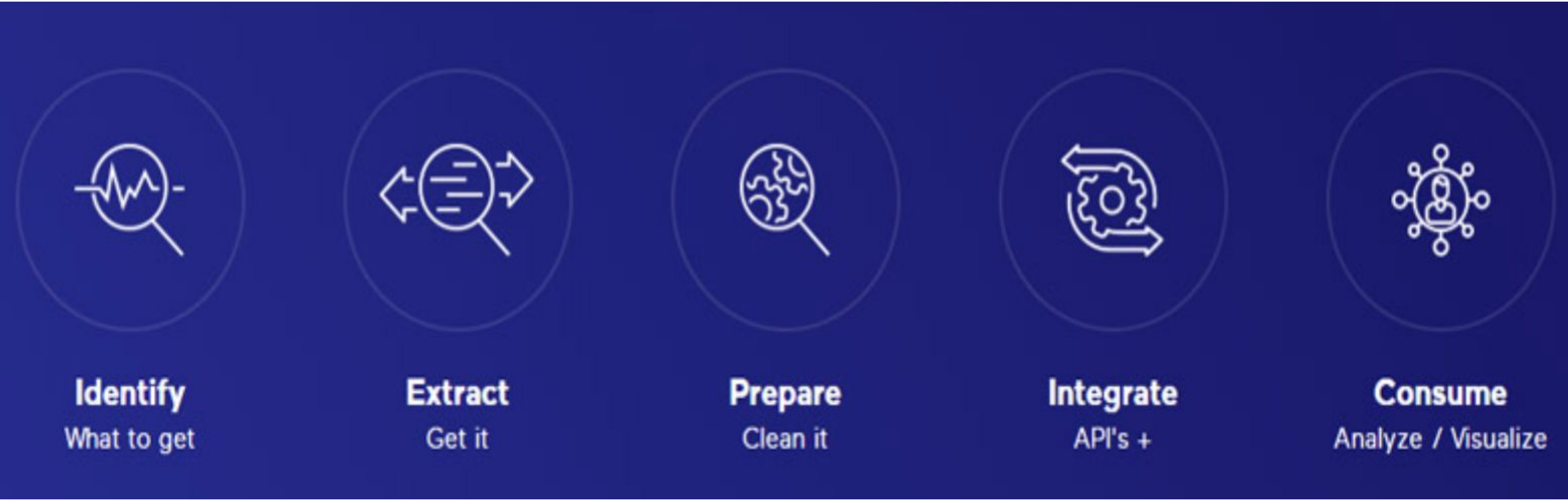
In simple words,

Data analysis is the process of **collecting** and **organizing data** in order to draw helpful conclusions from it.

Data Analysis Methods

- **Qualitative Analysis:** This approach mainly answers questions such as ‘why,’ ‘what’ or ‘how.’
- **Quantitative Analysis:** Generally, this analysis is measured in terms of numbers. The data here present themselves in terms of measurement scales and extend themselves for more statistical manipulation.

Data Analysis Process



Identify
What to get

Extract
Get it

Prepare
Clean it

Integrate
API's +

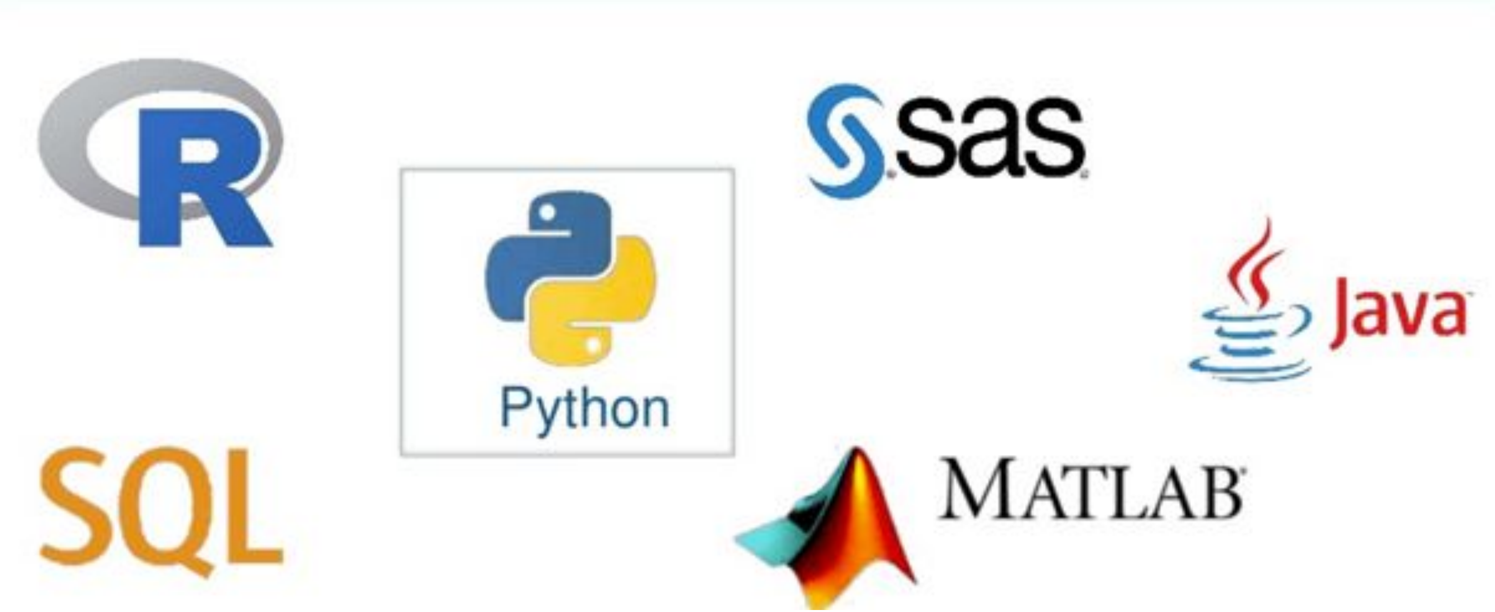
Consume
Analyze / Visualize

Python For Data Analysis

Why Data Analysis?

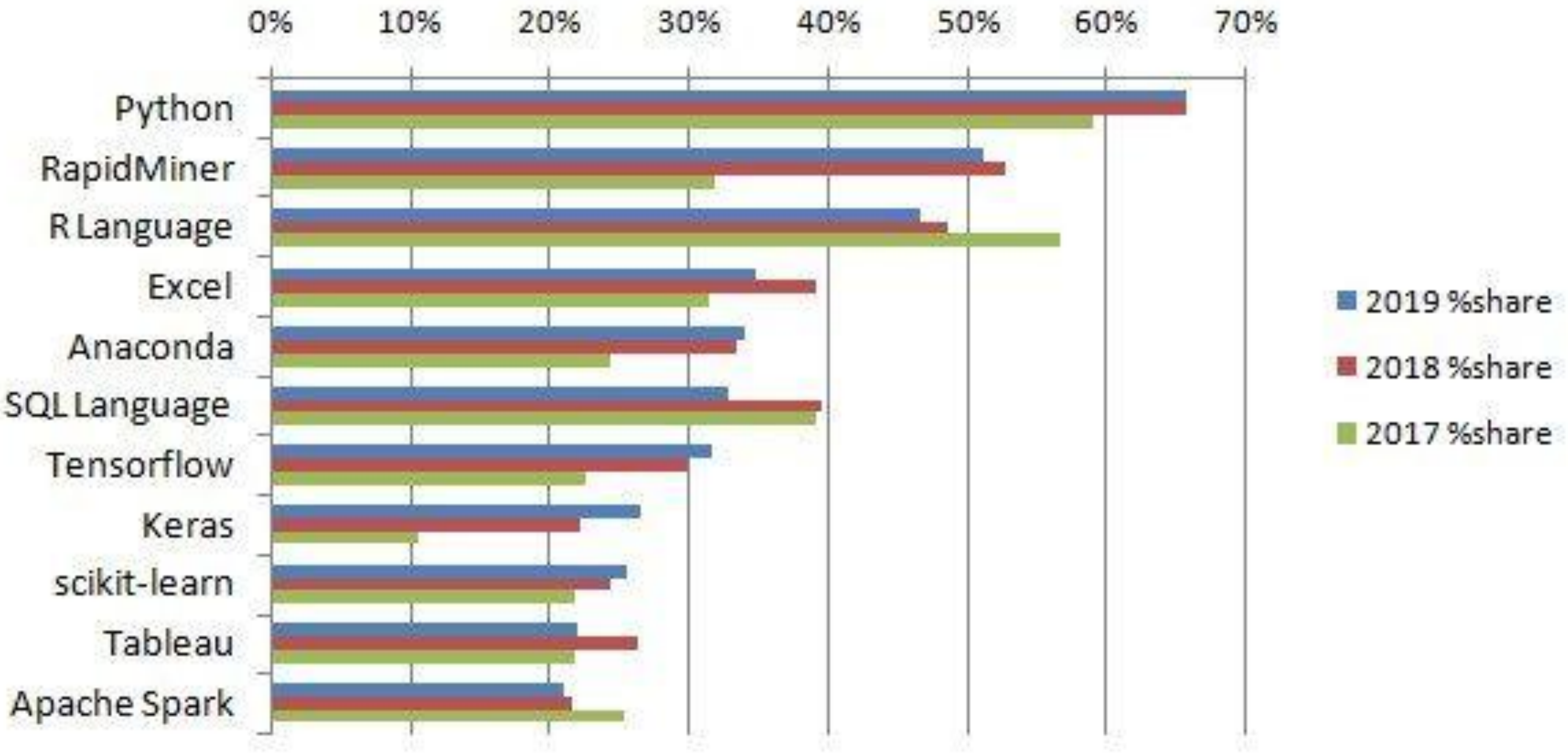
- Better Targeting
- New Innovations
- Cut Costs of Operation
- Helps Solve Problems

Data Analysis Tools



Data Analysis

Top Analytics, Data Science, Machine Learning Software 2017-2019, KDnuggets Poll



```
a=3
b=5
Sum=a+b
```

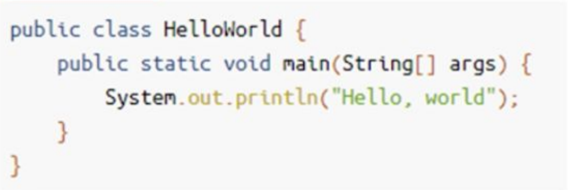
Open Source

High-level

Interpreted

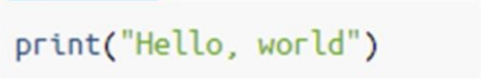
Large community

Java

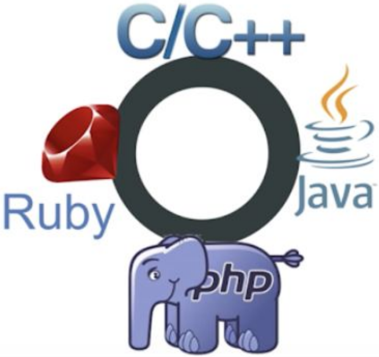


```
public class HelloWorld {
    public static void main(String[] args) {
        System.out.println("Hello, world");
    }
}
```

Python



```
print("Hello, world")
```



It’s that **SIMPLE**!

python

- Well-suited for data manipulation & analysis
- Deals with tabular data with heterogeneously-typed columns
- Arbitrary matrix data
- Observational/ statistical datasets

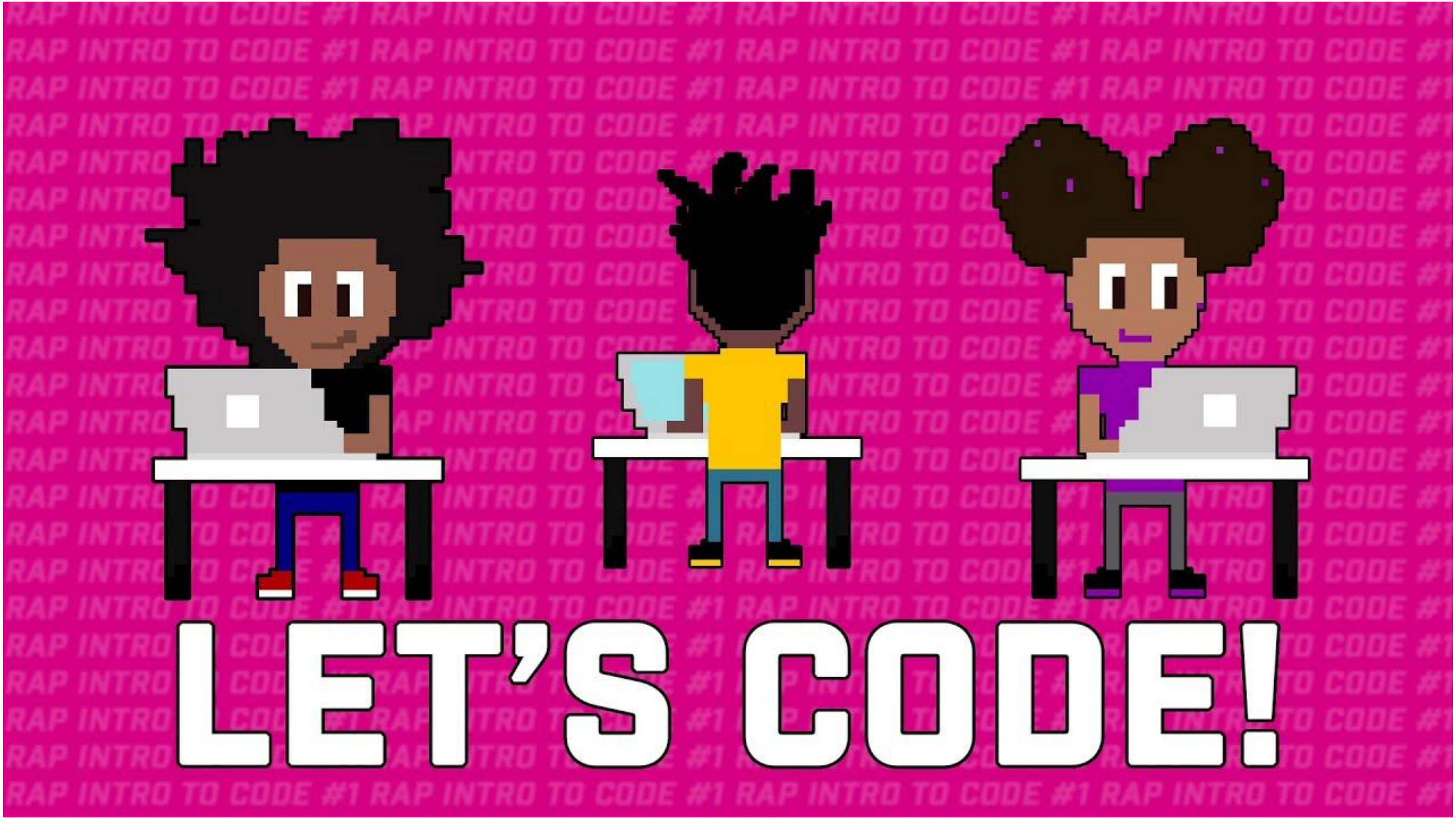
Libraries

NumPy | Pandas | matplotlib | seaborn

Jupyter Notebook	Google Colab
● Open source web application which is maintained by the people at Project Jupyter.	● Colaboratory is a free Jupyter notebook environment offered and maintained by Google
● You have to pip install the libraries	● Colab comes with libraries pre installed (you need not pip most of the libraries)
● It uses the local Machine’s kernel	● Google Colab runs on Google Cloud Platform (GCP). Hence it’s robust, flexible
● Jupyter Notebooks store the ipython notebooks locally in the Machine	● Google Colab comes with collaboration backed in the product, everythong is stored in Google drive, which makes sharing and collaborating more efficient

Modules

- Overview of the basics of Python
- Python Data Structures
- Data Analysis Libraries
    1. Numpy
    2. Pandas
    3. Matplotlib


RAP INTRO TO CODE #1
LET'S CODE!

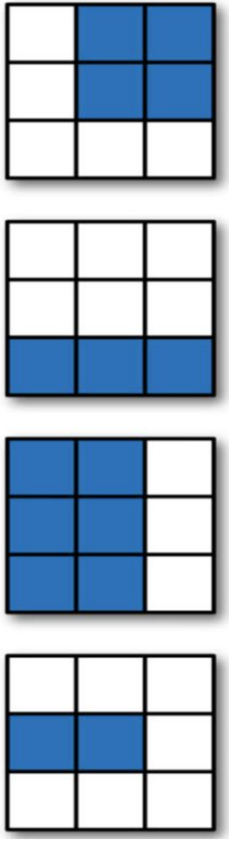
Python Collections

- List is a collection which is ordered and changeable. Allows duplicate members.
- Tuple is a collection which is ordered and unchangeable. Allows duplicate members.
- Set is a collection which is unordered and unindexed. No duplicate members.
- Dictionary is a collection which is unordered, changeable and indexed. No duplicate members.

Numpy

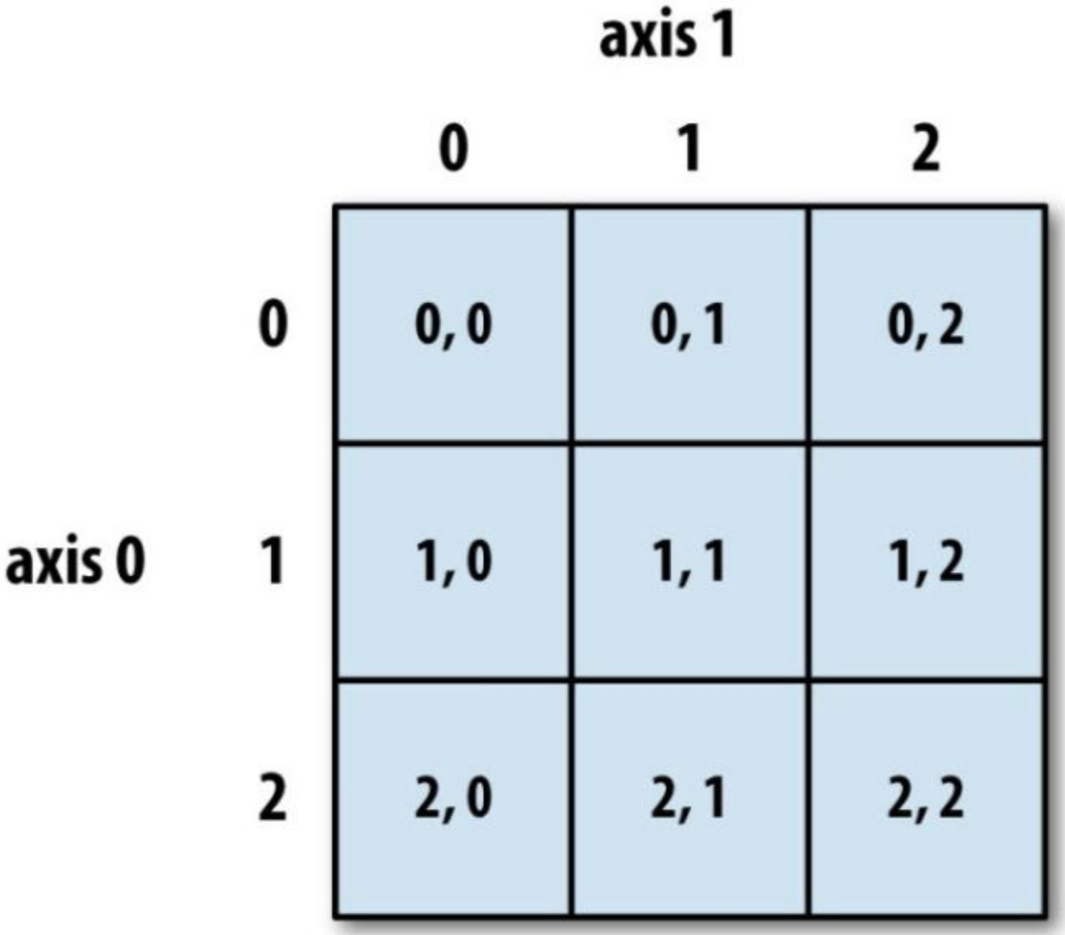
- NumPy is a python library used for working with arrays.
- It also has functions for working in domain of linear algebra, fourier transform, and matrices.
- NumPy was created in 2005 by Travis Oliphant. It is an open source project and you can use it freely.
- NumPy stands for Numerical Python.

Numpy Slicing



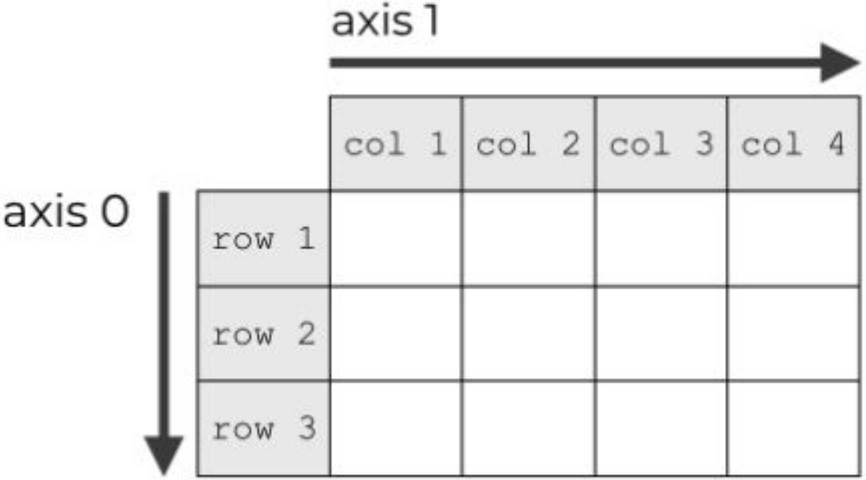
Expression	Shape
arr[:2, 1:]	(2, 2)
arr[2]	(3,)
arr[2, :]	(3,)
arr[2:, :]	(1, 3)
arr[:, :2]	(3, 2)
arr[1, :2]	(2,)
arr[1:2, :2]	(1, 2)

Axis



axis 1

		0	1	2
	0	0, 0	0, 1	0, 2
axis 0	1	1, 0	1, 1	1, 2
	2	2, 0	2, 1	2, 2

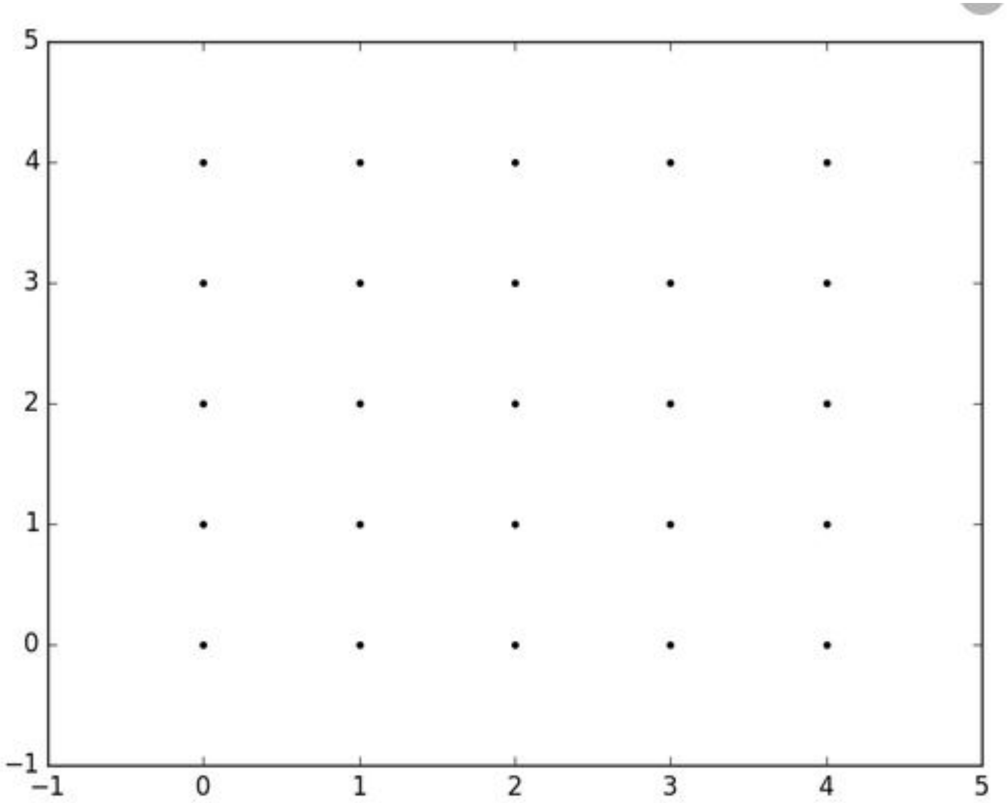


axis 1

axis 0

	col 1	col 2	col 3	col 4
row 1				
row 2				
row 3				

MeshGrid



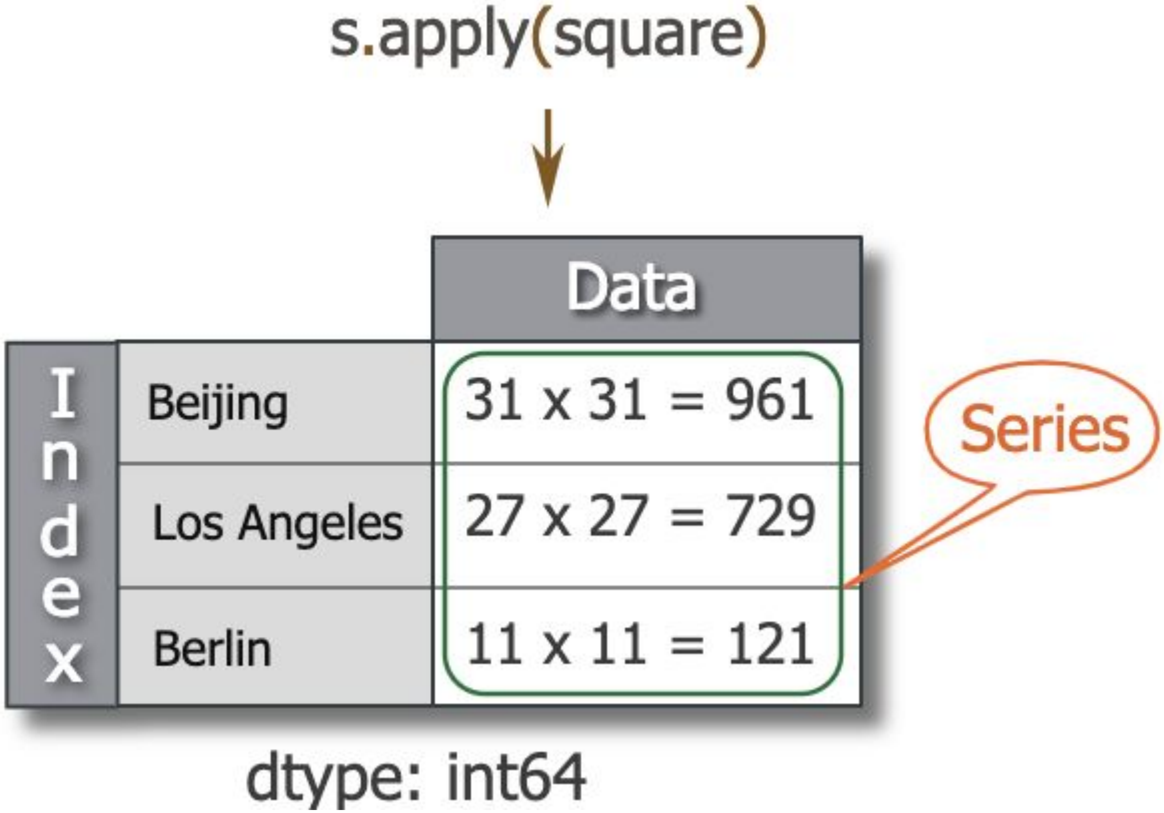
Pandas

Pandas is an open-source, BSD-licensed Python library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language

Majorly Two Data Types

1. Series
2. DataFrame

Series

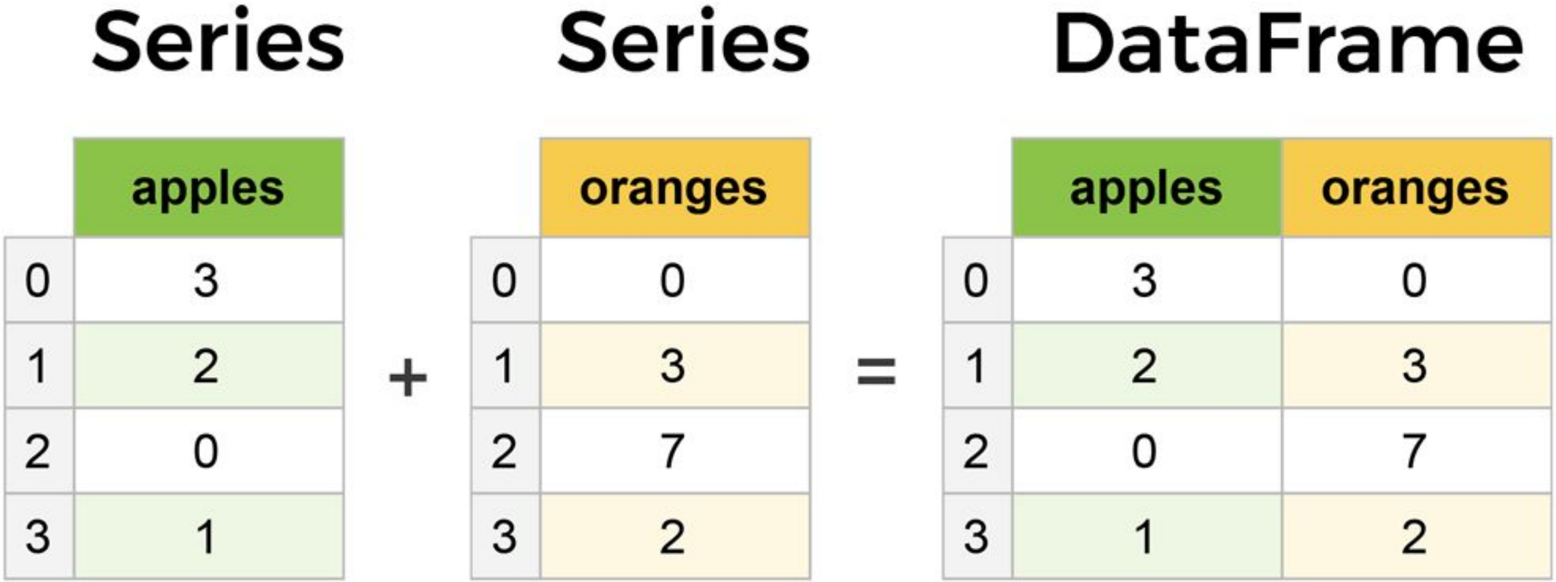


s.apply(square)

		Data
Index	Beijing	31 x 31 = 961
	Los Angeles	27 x 27 = 729
	Berlin	11 x 11 = 121

dtype: int64

DataFrame



Series

	apples
0	3
1	2
2	0
3	1

+

Series

	oranges
0	0
1	3
2	7
3	2

=

DataFrame

	apples	oranges
0	3	0
1	2	3
2	0	7
3	1	2

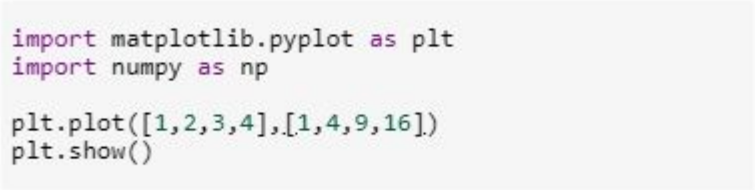
python

Intro

- A very powerful plotting library
- The most used module of Matplotlib is Pyplot
- Uses Python and it is open source.

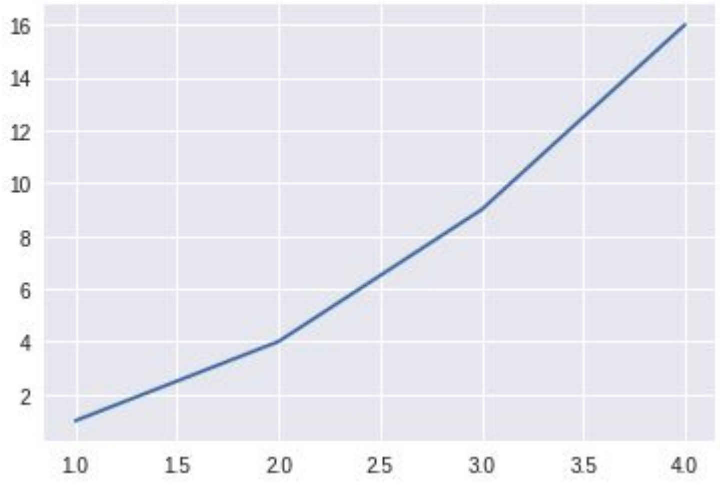
First Plot

- We pass two arrays as our input arguments to pyplot’s `plot()` method use `show()` method to invoke the required plot.
- The first array appears on the x-axis and second array appears on the y-axis of the plot.
- We add the title, and name x-axis and y-axis using methods `title()`, `xlabel()` and `ylabel()` respectively.



```
import matplotlib.pyplot as plt
import numpy as np

plt.plot([1,2,3,4],[1,4,9,16])
plt.show()
```

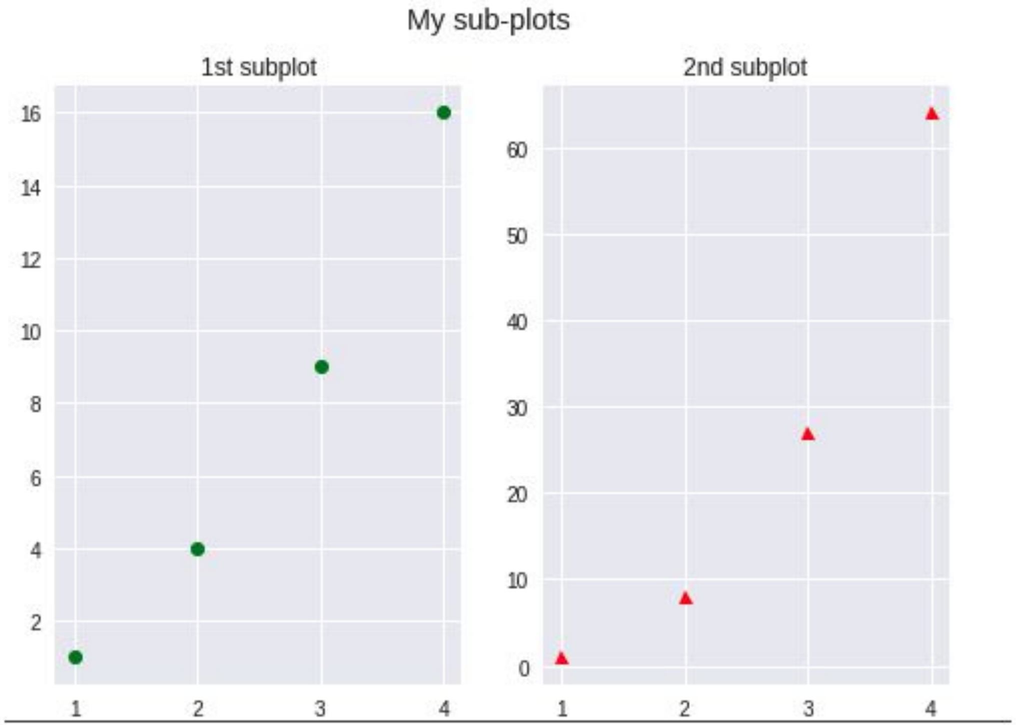


Multiple plots in one figure:

- **`subplot()`** method to add more than one plots in one figure

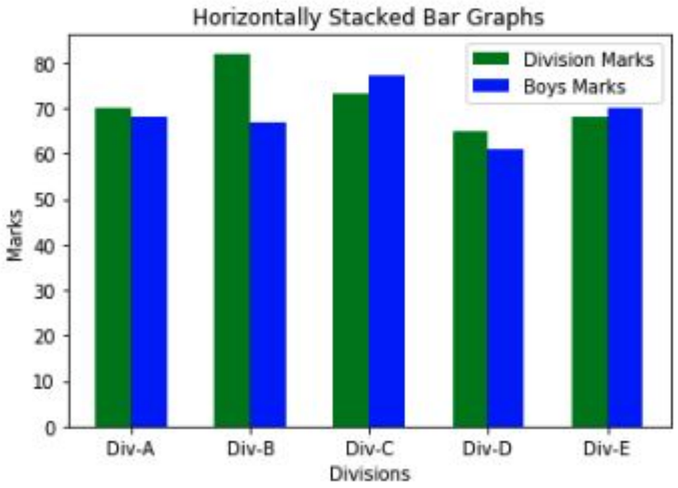
- The **`subplot()`** method takes three arguments: they are `nrows`, `ncols` and `index`. (1,2,1), (1,2,2)

- They indicate the number of rows, number of columns and the index number of the sub-plot.


My sub-plots
1st subplot
16
14
12
10
8
6
4
2
1
2
3
4
2nd subplot
60
50
40
30
20
10
0
1
2
3
4

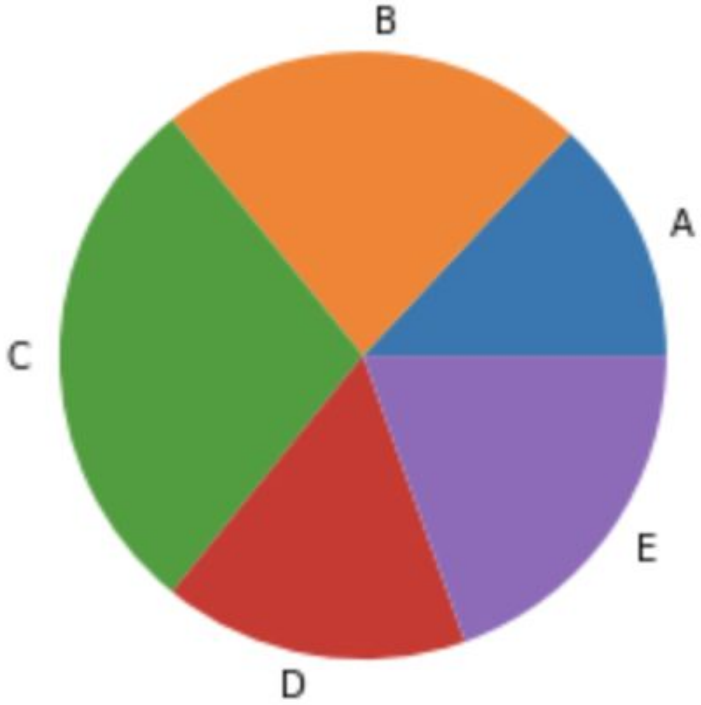
Stacked Plots

To create horizontally stacked bar graphs we use the `bar()` method twice and pass the arguments where we mention the **index and width** of our bar graphs in order to horizontally stack them together.



Pie Charts

One more basic type of chart is a Pie chart which can be made using the method `pie()` We can also pass in arguments to customize our Pie chart to show shadow, explode a part of it etc.



Histograms

- Data like height and weight, stock prices, waiting time for a customer, etc which are continuous in nature.

- Range against its frequency

- Probability and statistics like the normal -distribution

Scatter Plot

- Used especially they come in handy in visualizing a problem of regression.

- Relation between Height-weight, length-breadth,etc.

It’s just the beginning

Myself =

{

‘ Name ’ : ‘Sujitkumar Singh’,

‘ ✉ ’ : ‘2017.sujitkumar.singh@ves.ac.in’ ,

‘ ’ : ‘ github.com/singhsujitkumar’ ,

‘ ’ : ’@suj.eat’ ,

‘ ’ : ‘@sujitsofficial’

}

Github repo: https://github.com/SinghSujitkumar/DataAnalysisWrokshop