

## ZETECH UNIVERSITY

### DATA SCIENCE PROGRAMMIG WITH PYTHON

#### Lesson 3: Python Libraries for data science. A Statistics use case with python libraries

- Python Libraries for data science.
- 

#### Data science and machine learning: An overview

**Data science** - *is a field of applied mathematics and statistics that provides useful information based on the analysis and modeling of large amounts of data.*

**Machine learning** *is a branch of artificial intelligence and computer science that involves developing computer systems that can learn and adapt using algorithms and statistical models.*

#### Some of the real-world applications of data science and machine learning include:

- Google has identified **breast cancer tumors** that metastasize to nearby lymph nodes using a machine-learning tool called LYNA. The tool identified metastatic cancer with 99% accuracy using its algorithm, but more testing is needed before doctors can use it.
- A company called **Streetlight is modeling traffic patterns** for cars, bikes, and pedestrians in North America using data science and trillions of data points from smartphones and in-vehicle navigation devices.
- UPS is **optimizing package transportation** with a platform called Network Planning Tools that uses artificial intelligence and machine learning to work around bad weather and service bottlenecks.
- **RSPCT's** shooting-analysis system for basketball transmits data from a sensor on the hoop's rim to a device that displays shot details and generates predictive insights. The system has been adopted by NBA and college teams.
- The IRS has improved its **fraud detection** with taxpayer profiles built from public social media data, assorted metadata, emailing analysis, and electronic payment patterns. Based on those profiles, the IRS forecasts individual tax

returns, and anyone whose returns diverge wildly gets flagged for auditing. (Privacy advocates have not been pleased.)

- A company called Sovrn created **intelligent advertising technology** compatible with Google and Amazon's server-to-server bidding platforms to broker deals between advertisers and outlets.

### Why Python is used by data scientists

Python is not the only language used in data science and machine learning. R is another dominant option, and **Java**, **JavaScript**, and **C++** also have their places. But Python's advantages have helped it earn its place as one of the most popular programming languages generally, and in data science and machine learning specifically.

### These advantages include:

- Python is relatively **easy to learn**. Its syntax is concise and resembles English, which helps make learning it more intuitive.
- It has a large community of users. This translates into excellent **peer support and documentation**.
- Python is **portable** and allows you to run its code anywhere. This means a Python application can run across Windows, MacOS, and Linux without modifications to its source code (unless there are system-specific calls).
- Python is **a free, open-source**, and **object-oriented programming** language.
- Python makes it **easy to add modules** from other languages, such as C and C++.
- Finally, **many of Python's libraries were literally made for data science and machine learning**.

## Python Libraries for Data Processing and Model Deployment

### 1) Pandas

data analysis that contains high-level data structures and tools to manipulate data in a simple way.

### Key Features of Pandas

#### i) The Series and DataFrame Objects

#### ii) Restructuring of Data Sets

Pandas python provides the flexibility for reshaping the data structures to be inserted in both rows and columns of tabular data.

### iii) Labelling

To allow automatic data alignment and indexing, pandas provide labeling on series and tabular data.

### v) Grouping

The functionality to perform split-apply-combine on series as well on tabular data.

### vi) Identify and Fix Missing Data

Programmers can quickly identify and mix missing data floating and non-floating pointing numbers using pandas.

### vii) Powerful capabilities to load and save data from various formats such as JSON, CSV, HDF5, etc.

## 2) NumPy

Below are some of the features provided by NumPy-

Integration with legacy languages.

Mathematical Operations: It provides all the standard functions required to perform operations on large data sets swiftly and efficiently, which otherwise have to be achieved through looping constructs.

ndarray: It is a fast and efficient multidimensional array that can perform vector-based arithmetic operations and has powerful broadcasting capabilities.

I/O Operations: It provides various tools which can be used to write/read huge data sets from disk. It also supports I/O operations on memory-based file mappings.

Fourier transform capabilities, **Linear Algebra, and Random Number Generation.**

## SciPy

Visualizing and manipulating data with high-level commands and classes.

Python sessions that are both robust and interactive.

For parallel programming, there are classes and web and database procedures.

## 4) Sci-Kit Learn

Sci-Kit Learn

The scikit learn library is a helpful tool to predict customer behavior, develop neuroimages, and more.

It's simple to use and completely free.

## 6) Tensorflow

TensorFlow is a free end-to-end open-source platform for Machine Learning that includes a wide range of tools, libraries, and resources. The Google Brain team first released it on November 9, 2015. TensorFlow makes it simple to design and train Machine Learning models using high-level APIs like Keras.

### Key Features of TensorFlow

It is a Google-developed open-source framework.

Deep learning networks and machine learning principles are supported.

It's simple to use and provides for rapid debugging.

## 7)Matplotlib

Matplotlib is a **data visualization library and 2-D plotting library** of Python. It was initially released in 2003 and it is the most popular and widely-used plotting library in the Python community.

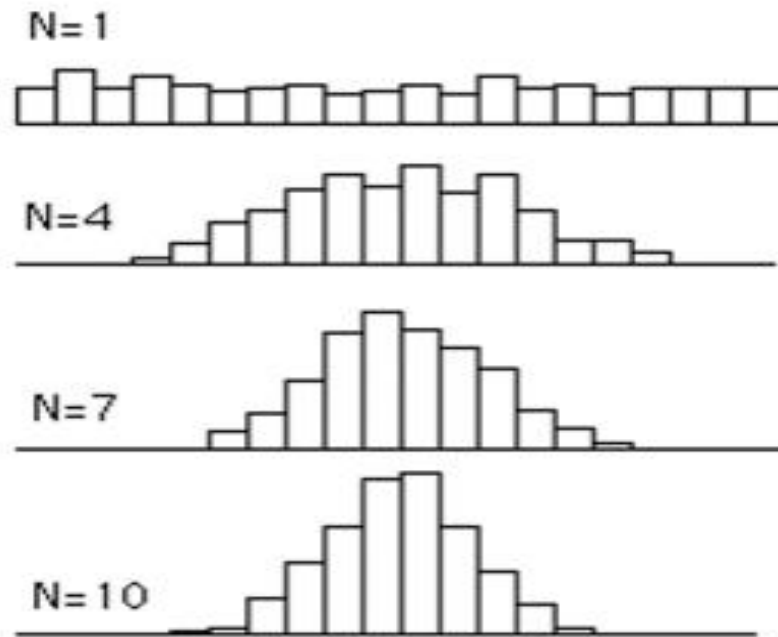
<https://www.dataquest.io/blog/15-python-libraries-for-data-science/>

### CENTRAL LIMIT THEOREM IN DATA SCIENCE

- ✚ CLT is **the critical component of the data science life cycle**, and a main core of the **hypothesis testing**.
- ✚ Central limit theorem
  - *describes the shape of the distribution of sample means as a Gaussian, which is a distribution that statistics knows a lot about.*
- ✚ The theorem states that as the size of the sample increases, the distribution of the mean across multiple samples will approximate a Gaussian distribution.

### Why learn Central Limit theorem

- **Central limit theorem** - *helps us to make inferences about the sample and population parameters and construct better machine learning models using them.*



When  $n$  increases:

1. the distributions become more and more normal.
2. the spread of the distributions decreases.

#### Statistical Significance

- Analyzing data involves statistical methods *like hypothesis testing and constructing confidence intervals*. These methods assume that the population is normally distributed. In the case of unknown or non-normal distributions, we treat the sampling distribution as normal according to the central limit theorem
- **If we increase the samples drawn from the population**, the standard deviation of sample means will decrease. *This helps us estimate the population mean much more accurately*
- Also, the sample mean can be used to create the range of values known as a confidence interval (that is likely to consist of the population mean)

#### Practical Application

As an example:

- **Political/election** - polls are prime CLT applications. These polls estimate the percentage of people who support a particular candidate. You might have seen these results on news channels that come with confidence intervals. The central limit theorem helps calculate that

- **Confidence interval**, - an application of CLT, is used to calculate the mean family income for a particular region.

The mean of the sample means is denoted as:

$$\mu_{\bar{X}} = \mu$$

where,

$\mu_{\bar{X}}$  = Mean of the sample means

$\mu$  = Population mean

And, the standard deviation of the sample mean is denoted as:


$$\sigma_{\bar{X}} = \sigma / \sqrt{n}$$

where,











$\sigma_{\bar{X}}$  = Standard deviation of the sample mean

$\sigma$  = Population standard deviation

$n$  = sample size


**jupyter**
 Untitled35
 Last Checkpoint: an hour ago (autosaved)

File Edit View Insert Cell Kernel Widgets Help











 Code

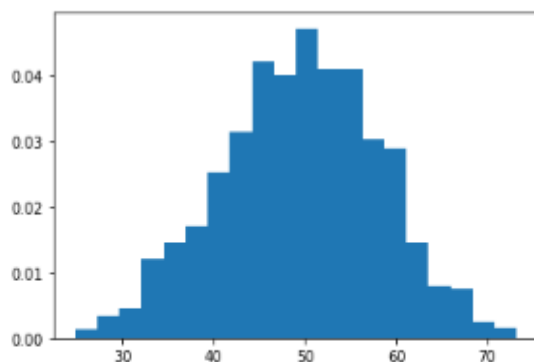
```
In [1]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

Matplotlib is building the font cache; this may take a moment.

```
In [2]: x=[]
```

```
In [3]: distribution = np.random.randint(0,100,100)
for i in range(1000):
    index = np.random.choice(distribution.shape[0], 10, replace=False)
    x.append(np.mean(index))
plt.hist(x,20, density=True)
```

```
Out[3]: (array([0.00123967, 0.00330579, 0.00454545, 0.01198347, 0.01446281,
0.01694215, 0.02520661, 0.03140496, 0.04214876, 0.04008264,
0.04710744, 0.04090909, 0.04090909, 0.03016529, 0.02892562,
0.01446281, 0.00785124, 0.00743802, 0.00247934, 0.00165289]),
array([24.8 , 27.22, 29.64, 32.06, 34.48, 36.9 , 39.32, 41.74, 44.16,
46.58, 49. , 51.42, 53.84, 56.26, 58.68, 61.1 , 63.52, 65.94,
68.36, 70.78, 73.2 ]),
<BarContainer object of 20 artists>)
```

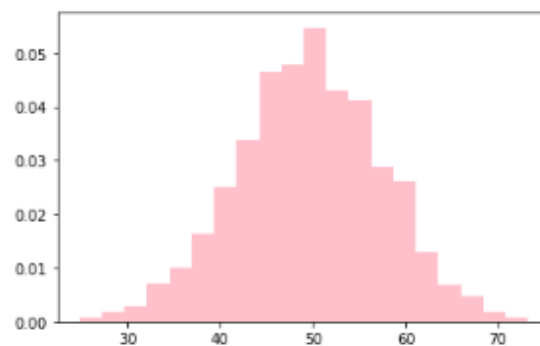


In the above code, I have created a random distribution with 100 values ranging from 0 to 100. '[Line-5]

```
In [4]: distribution = np.random.randint(0,100,100)
for i in range(1000):
    index = np.random.choice(distribution.shape[0], 15, replace=False)
    x.append(np.mean(index))
plt.hist(x,20, density=True, color='pink')
sns.kdeplot(x)

-----
NameError                                Traceback (most recent call last)
Input In [4], in <cell line: 6>()
      4     x.append(np.mean(index))
      5 plt.hist(x,20, density=True, color='pink')
----> 6 sns.kdeplot(x)

NameError: name 'sns' is not defined
```



In [ ]:

- **density:** This parameter is an optional parameter and it contains the boolean values.
- **weights:** This parameter is an optional parameter and it is an array of weights, of the same shape as x
- **label :** This parameter is an optional parameter and it is a string, or sequence of strings to match multiple datasets.
- **bins :** This parameter is an optional parameter and it contains the integer or sequence or string.
-