# A SEMINAR REPORT ON

*MOST COMMON LIBRARIES FOR DATA SCIENCE*

*Submitted in partial fulfillment of the requirements for the award of the degree of*

# BACHELOR OF TECHNOLOGY

In

# **COMPUTER SCIENCE AND ENGINEERING** C:\Users\VENKATESWARLU\Desktop\khit logo.webp

*Submitted by*

M.HARSHINI

 178X1A0563

**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**

**KALLAM HARANADHAREDDY INSTITUTE OF TECHNOLOGY AN ISO 9001:2015 CERTIFIED INSTITUTION**

**ACCREDITED BY NBA & NAAC WITH ‘A’ GRADE (APPROVED BY AICTE, AFFILIATED TO JNTUK, KAKINADA) NH-5, CHOWDAVARAM, GUNTUR – 522019**

**MARCH - APRIL 2021**

# **KALLAM HARANADHAREDDY INSTITUTE OF TECHNOLOGY**

**AN ISO 9001:2015 CERTIFIED INSTITUTION ACCREDITED BY NBA & NAAC WITH ‘A’ GRADE**

**(APPROVED BY AICTE, AFFILIATED TO JNTUK, KAKINADA) NH-5, CHOWDAVARAM, GUNTUR-522019**

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

**CERTIFICATE**

This is to certify that the seminar entitled “**MOST COMMON LIBRARIRS FOR DATA SCIENCE**” being submitted by M.HARSHINI (178X1A0563) in the partial fulfilment for the award of degree of Bachelor of Technology in Computer Science and Engineering in Kallam Haranadhareddy Institute of Technology and this bonafide work carried out by him/her.

**Internal Examiner Head of Department**

**Mr. S. SRINIVASA RAO, Dr.k.VENKATASUBBAREDDY**

**Assistant Professor B. Tech., M. Tech. professor & HOD M. Tech, Ph. D.**

###### TABLE OF CONTENTS

###### **1. abstract. 1**

###### **2. USED LIBRARIES. 2**

###### **3. ARCHITECTURE. 3**

**4. NUMPY. 4 to 5**

**5. PANDAS 6 to 7**

**6. MATPLOTLIB. 8 to 9**

**7. SCIPY. 10 to 11**

**8. TENSOR FLOW. 12 to 13**

**9. ADVANTAGES OF DATA SCIENCE 14**

**10. DISADVANTAGES OF DATA SCIENCED 15**

**11. CONCLUSION 16**

**12. REFERENCE 17**

**ABSTRAC****T**

Python is the most widely used programming language today. When it comes to solving data science tasks and challenges, Python never ceases to surprise its users. Most [data scientists](https://www.simplilearn.com/data-science-career-guide-pdf" \t "_blank) are already leveraging the power of Python programming every day. [Python](https://www.simplilearn.com/tutorials/python-tutorial" \t "_blank) is an easy-to-learn, easy-to-debug, widely used, object-oriented, open-source, high-performance language, and there are many more benefits to Python programming. Python has been built with extraordinary Python libraries for data science that are used by programmers every day in solving problems. Here’s the top 10 Python libraries for data science:

###### **USED LIBRARIES**

###### .

**. NumPy**

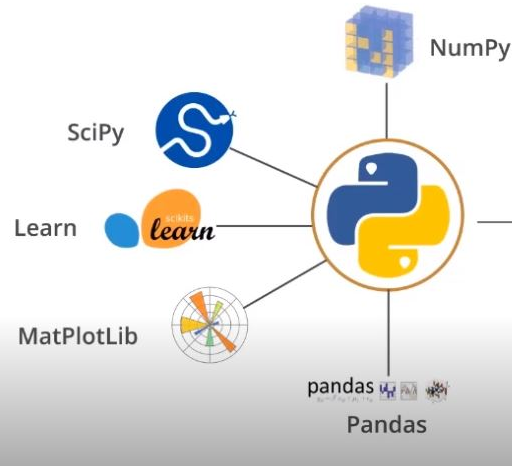
**. Pandas**

**. SciPy**

**. Matplotlib**

**. TensorFlow**

**ARCHITECTURE**



\* EASY TO LEARN

\* OPEN SOURCE

\* BIG OPEN SOURCE COMMUNITY

\* EFFICIENT AND MULTI PLATFORM SUPPORT

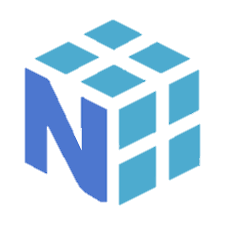
\* INTEGRATES WELL WITH ENTERPRIES APPS AND SYSTEMS

\* GREAT VENDOR AND PRODUCT SUPPORT

\* HUGE COLLECTION OF LIBRARIES,FUNCTIONS AND MODULE

###### NumPy

 NUMPY (Numerical Python) is the fundamental package for numerical computation in Python; it contains a powerful N-dimensional array object. It’s a general-purpose array-processing package that provides high-performance multidimensional objects called arrays and tools for working with them.

 NumPy also addresses the slowness problem partly by providing these multidimensional arrays as well as providing functions and operators that operate efficiently on these arrays. It has around 18,000 comments on GitHub and an active community of 700 contributors.

**Features:**

Provides fast, precompiled functions for numerical routines

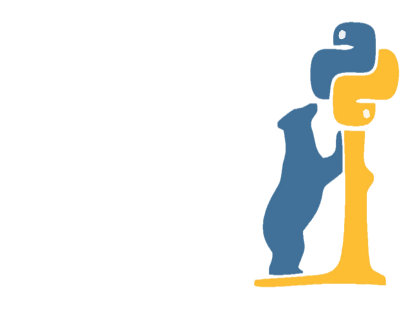
* Array-oriented computing for better efficiency
* Supports an object-oriented approach
* Compact and faster computations with vectorization

### **Applications:**

* Extensively used in data analysis
* Creates powerful N-dimensional array
* Forms the base of other libraries, such as SciPy and scikit-learn
* Replacement of MATLAB when used with SciPy and matplotlib

**PANDAS:**

PANDAS (Python data analysis) is a must in the data science life cycle. It is the most popular and widely used Python library for data science, along with NumPy in matplotlib. With around 17,00 comments on GitHub and anactive community of 1,200 contributors

 .

it is heavily used for [data analysis](https://www.simplilearn.com/data-analysis-methods-process-types-article" \t "_blank) and cleaning. Pandas provides fast, flexible [data structures](https://www.simplilearn.com/tutorials/data-structure-tutorial" \t "_blank), such as data frame CDs, which are designed to work with structured data very easily and intuitively.

**Features:**

* Eloquent syntax and rich functionalities that gives you the freedom to deal with missing data
* Enables you to create your own function and run it across a series of data
* High-level abstraction
* Contains high-level data structures and manipulation tools

**Applications:**

* General [data wrangling](https://www.simplilearn.com/data-wrangling-article" \t "_blank) and [data cleaning](https://www.simplilearn.com/data-cleaning-why-and-how-to-get-started-article" \t "_blank)

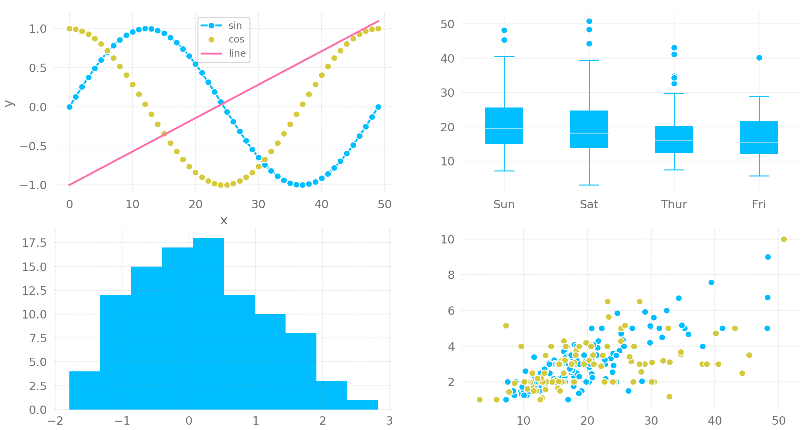
ETL (extract, transform, load) jobs for data transformation and data storage, as it has excellent support for loading CSV files into its data frame format

* Used in a variety of academic and commercial areas, including statistics, finance and neuroscience
* Time-series-specific functionality, such as date range generation, moving window, linear regression and date shifting.

###### 

###### **Matplotlib**

MATPLOTLIB  has powerful yet beautiful visualizations. It’s a plotting library for Python with around 26,000 comments on GitHub and a very vibrant community of about 700 contributors. Because of the graphs and plots that it produces, it’s extensively used for data visualization. It also provides an object-oriented API, which can be used to embed those plots into applications.



### **Features:**

* Usable as a MATLAB replacement, with the advantage of being free and open source
* Supports dozens of backends and output types, which means you can use it regardless of which operating system you’re using or which output format you wish to use
* Pandas itself can be used as wrappers around MATLAB API to drive MATLAB like a cleaner
* Low memory consumption and better runtime behavior

### **Applications:**

* Correlation analysis of variables
* Visualize 95 percent confidence intervals of the models
* Outlier detection using a scatter plot etc.
* Visualize the distribution of data to gain instant insights

SCIPY

SciPy (Scientific Python) is another free and open-source Python library for data science that is extensively used for high-level computations. SciPy has around 19,000 comments on GitHub and an active community of about 600 contributors. It’s extensively used for scientific and technical computations, because it extends NumPy and provides many user-friendly and efficient routines for scientific calculations.



### **Features:**

* collection of algorithms and functions built on the NumPy extension of Python
* High-level commands for data manipulation and visualization
* Multidimensional image processing with the SciPy ndimage submodule
* Includes built-in functions for solving differential equations

### **Applications:**

* Multidimensional image operations
* Solving differential equations and the Fourier transform
* Optimization algorithms
* Linear algebra

## **TENSORFlOW**

The first in the list of python libraries for data science is TensorFlow. [TensorFlow](https://www.simplilearn.com/tutorials/deep-learning-tutorial/tensorflow" \t "_blank) is a library for high-performance numerical computations with around 35,000 comments and a vibrant community of around 1,500 contributors. It’s used across various scientific fields. TensorFlow is basically a framework for defining and running computations that involve tensors, which are partially defined computational objects that eventually produce a value.



### **Features:**

* Better computational graph visualizations
* Reduces error by 50 to 60 percent in neural machine learning
* Parallel computing to execute complex models
* Seamless library management backed by Google
* Quicker updates and frequent new releases to provide you with the latest features

**Applications:**

* Speech and image recognition
* Text-based applications
* Time-series analysis
* Video detection

ADVANTAGES:

\* EXTENSIVE LIBRARIES

\* SIMPLE ANSD EASY

\* POTABLE

\* IMPROVED PRODUCTIVITY

\* EMBEDDABLE

\* IOT OPPORTUNITIES

DISADVANTAGES:

\* SPEED LIMITATION

\* WEAK IN MOBILE COMPUTING AND BROWSERS

\* DESIGN RESTRICTION

\* SIMPLE

\* UNDER DEVELOPED DATABASE ACCESS LAYERS

###### **Conclusion**

###### While The Python Language Reference describes the exact syntax and semantics of the Python language, this library reference manual describes the standard library that is distributed with Python. It also describes some of the optional components that are commonly included in Python distributions.

###### Python’s standard library is very extensive, offering a wide range of facilities as indicated by the long table of contents listed below. The library contains built-in modules (written in C) that provide access to system functionality such as file I/O that would otherwise be inaccessible to Python programmers, as well as modules written in Python that provide standardized solutions for many problems that occur in everyday programming. Some of these modules are explicitly designed to encourage and enhance the portability of Python programs by abstracting away platform-specifics into platform-neutral APIs.

###### **REFERENCES**

###### **• Guttag, John V. (12 August 2016). Introduction to Computation and Programming Using Python: With Application to Understanding Data. MIT Press. ISBN 978-0-262-52962-4.**

###### **• "Python Release Python 3.8.2". Python.org.**

###### **• "Python Release Python 3.9.0a5". Python.org.**

###### **• Benjamin (6 April 2020). "Python Insider: Python 2.7.18 release candidate 1 available". Python Insider. Retrieved 14 April 2020. Python 2.7.18 will be the last release of the Python 2.7 series, and thus Python 2.**