BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Python for Data Science

Table of Contents

Course Objectives

CO1	Introduce students with fundamental programming concepts of Python
CO2	Enable students to solve data problems using Python

Text and References

T1	Charles Severance: Python for Everybody, Exploring Data in Python 3, Creative Commons, 2016
T2	Jake VanderPlas: Python Data Science Handbook, Essential Tools for Working with Data, O'Reilly Media, 2016
T3	Edouard Duchesnay: Statistics and Machine Learning in Python Release 0.2, 2018
T4	Wes McKinney: Python for Data Analysis, Agile Tools for Real World Data, O'Reilly Media, 2013

Additional Reading

1. **Python 3.* documentation**
2. **Numpy Documentation**
3. **Pandas Documentation**
4. **Matplotlib documentation**
5. **seaborn: statistical data visualization documentation**
6. **Scikit-learn documentation**

Modular Content Structure

Session	Topics	Reference
	Saturday, April 17, 2021 – S1	
1	**Python Basics**	
1.1	**Setting up Python Environments**	**Python Documentation**
1.1_1	Anaconda Distribution	
1.1_2	Spyder IDE	
1.1_3	Jupyter Notebooks	
1.1_4	Input / Output with Python	
1.2	**Getting familiarity with basic code constructs**	**T1 : Ch 2, Class Notes**
1.2_1	Package imports	
1.2_2	Data Types & Type Casting	
1.2_3	Variables, Expressions & Statements	
1.2_4	Object Orientation (Introduction)	
	Saturday, April 17, 2021 – S2	
2	**Python Data Structures**	
1.3	**Immutable Data Structures**	**T1 : Ch 6, 10, Class Notes**
1.3_1	Immutable Data Structures	
1.3_2	Strings	
1.3_3	Operations on String	
1.3_4	Familiarity with Tuples	
	Python Data Structures	
2.1	**Mutable Data Structures**	**T1 : Ch 8, 9, Class Notes**
2.1_1	List	
2.1_2	List operations	
2.1_3	Familiarity with Sets	
2.1_4	Dictionary operations	
	Python Programming Constructs	
2.2	**Expressions, Operations, and Decision Structures**	**T1 : Ch 2, 3, Class Notes**
2.2_1	Boolean Expressions and Logical Operators	
2.2_2	Conditional and Alternative execution	
2.2_3	Chained and Nested execution	
2.2_4	Catching Exceptions with try and except	
	Sunday, April 18, 2021 – S1	
2.3	**Iterative Executions**	**T1 : Ch 5, Class Notes**
2.3_1	While loops	
2.3_2	Infinite loops, break, continue	
2.3_3	For loops	
2.3_4	Loop patterns	

3	**Functions and Files**	
3.1	**Functions**	**T1 : Ch 4, Class Notes**
3.1_1	Functions calls	
3.1_2	Built in Functions	
3.1_3	Custom Functions	
3.1_4	Parameters and Arguments	
3.2	**Files**	**T1 : Ch 7, Class Notes**
3.2_1	Opening files	
3.2_2	Reading files	
3.2_3	Operation on content of files	
3.2_4	Writing files	
	Saturday, April 24, 2021 – S1	
3.3	**SciPy Ecosystem**	**SciPy Documentation**
3.3_1	Familiarity with SciPy Ecosystem	
3.3_2	NumPy Library	
3.3_3	SciPy Library	
3.3_4	Matplotlib Library	
4	**NumPy**	
4.1	**Multidimensional Arrays with NumPy**	**T2 : Ch 2, Class Notes**
4.1_1	Basics of NumPy Arrays	
4.1_2	Computation on NumPy Arrays	
4.1_3	Aggregations	
4.1_4	Structured Arrays	
	Data Transformation with Pandas	
4.2	**Data Exploration I**	**Pandas Documentation**
4.2_1	Pandas Objects	
4.2_2	Data Indexing and Selection	
4.2_3	Reading files with Pandas	
4.2_4	Dataset Merges	
	Saturday, April 24, 2021 – S2	
4.3	**Data Exploration II**	**T2 : Ch 3, Class Notes**
4.3_1	Data Cleaning	
4.3_2	Data Transformation	
4.3_3	Data Filtering	
4.3_4	Aggregation and Grouping	
5	**Data Visualizations**	
5.1	**Visualizations with MatplotLib**	**Documentation, Class Notes**
5.1_1	Basic Plotting	
5.1_2	Life cycle of a Plot	
5.1_3	Subplots	
5.1_4	Plotting visuals	

	Saturday, April 25, 2021 – S1	
5.2	**Visualizations with Seaborn**	**Documentation, Class Notes**
5.2_1 5.2_2 5.2_3 5.2_4	Visualizing statistical relations Plots for univariate and multivariate analysis Visualizing distributions Linear relationships with plots	
	Introduction - Machine Learning with Python	
5.3	**Basic Machine Learning Examples with Python**	
5.3_1 5.3_2 5.3_3 5.3_4	Introducing Machine Learning Familiarity with Scikit-learn library Linear Regression - Handcoding Linear Regerssion − with Scikit-learn	**Scikit-learn documentation** **T2 : Ch 5, Class Notes**

Sample Labs / Assignments

1) Implement following custom string methods

aString_count ()	Returns the number of times a specified value occurs in a string
aString_count (string, value, start, end)	string Required. A string in which the other string to be searched value Required. A String. The string to value to search for start Optional. An Integer. The position to start the search. Default is 0 end Optional. An Integer. The position to end the search. Default is end of the string

aString_index ()	Searches the string for a specified value and returns the position of where it was found finds the first occurrence of the specified value method raises an exception if the value is not found
aString_index (string, value, start, end)	string Required. A string in which the other strings position to be determined value Required. The value to search for start Optional. Where to start the search. Default is 0 end Optional. Where to end the search. Default is to the end of the string

aString_isdigit ()	Returns True if all characters in the string are digits Returns True if all the characters are digits, otherwise False
aString_isdigit (string)	string Required. A string which needs to be determined as one with digits or without digits

2) An organization has data set with attributes like R&D cost, Administration and marketing cost for several States. Also, they have the record of the profit earned during a cycle. Implement a multiple linear regression model that will help organization to predict the profit in any cycle provided the costs and state are known for it. Refer the sample "Companies" dataset provided for this purpose.

3) Iris is the default dataset available in Scikit-learn library. Implement k-Nearest Neighbours (kNN) classifier with your custom code. You can consider Eucledian distance as the distance measure for the same. Also implement the same classifier with Scikit-learn. Compare the accuracy rates of both classifiers.