**Applied Data Science using Python Libraries like Pandas, Matplotlib and Scikit-Learn**

**Duration:** 3 days

**Pre-requisite:** Basic Python Programming

**Short Description:**

This Applied Data Science using Python course will provide you with a thorough understanding of each of the key Python libraries used for data science. Specifically, you will learn NumPy, Pandas, Matplotlib and Scikit-learn, known as the Python data stack. We will perform data exploration, analysis, visualization and modeling. The course will culminate with you applying these tools to a hands-on data science project.

**Long Description:**

This Applied Data Science using Python course will provide you with a thorough understanding of each of the key Python libraries used for data science. Specifically, you will learn NumPy, Pandas, Matplotlib and Scikit-learn, known as the Python data stack. We will perform data exploration, analysis, visualization and modeling. The course will culminate with you applying these tools to a hands-on data science project.

We will begin by discussing the data science process and how to effectively work through a data science problem. We’ll talk about how to clean, transform and prepare data for analysis. We will also cover descriptive and inferential statistics which will enable you to perform hypothesis testing so that you can better interpret the significance of your analysis. We will also focus on machine learning and predictive analytics. We’ll discuss the various ways to measure model performance, how to select the best model for your project and ways to refine that model.

**Learning Objectives:**

After this course, you will be able to:

* Install Anaconda on a personal computer.
* Have a clear understanding of data science and its role
* Understand the data science process
* Understand foundational descriptive statistics
* Understand foundational inferential statistics
* Understand the reasons for Python’s popularity in data science.
* Learn the primary libraries for data science in Python including NumPy, Pandas, Matplotlib and Scikit-learn.
* Interact with and manipulate data arrays and matrices using NumPy
* Perform exploratory data analysis using Pandas.
* Use Matplotlib and Seaborn to perform data visualization.
* Properly clean and prepare data for machine learning
* Apply machine learning on a variety of datasets
* Complete a data science project, end to end
* Understand the big picture and the importance of data science in industry, research and technology.

**Topic Outline:**

**Day 1**

* Course introduction
* Install Anaconda
* Overview of Data Science
* The data science process
* Identifying a problem and asking good questions
* Descriptive statistics
* **Milestone 1**: Learn how to use Jupyter Notebooks
* Essential libraries
  + Numpy
  + Pandas
  + Matplotlib
* **Milestone 2**: Exploratory data analysis

**Day 2**

* Getting data
* Feature selection
* Strategies for imputing missing data
* Inferential statistics
* Essential libraries
  + Statsmodels
  + Scikit-learn
* Confidence intervals
* Hypothesis testing
* **Milestone 3**: Significance testing
* Transforming data
  + Binary encoding
  + One-hot encoding
* Feature Engineering
* Training and test sets
* Standardizing data
* **Milestone 4**: Data modeling

**Day 3**

* Machine learning
* K-fold cross validation
* Box plot
* Measuring performance
* **Milestone 5**: Model selection
* Refining the model
* Hyperparameter tuning
* Grid search
* **Milestone 6**: End-to-end project
* Next steps

**Structured Activity/Exercises/Case Studies:**

Day 1:

* Milestone 1 – Learn how to use Jupyter Notebooks
* Milestone 2 – Exploratory data analysis

Day 2:

* Milestone 3 – Significance testing
* Milestone 4 – Data modeling

Day 3:

* Milestone 5 – Model selection
* Milestone 6 – End-to-end project

**Training material provided:** Yes (Digital format)