



Python for Data Analysis

Course ISI-1559 Five Days Instructor-led Hands on

Introduction

Data analysts are in demand everywhere today! This five-day, instructor-led course shows you how to do data analysis the way the pros do. You'll master descriptive analysis, using Pandas to analyze the data and Seaborn to create the visualizations that let you present your findings effectively. You'll get started with predictive analysis, using Scikit-learn with linear regression models. And you'll be guided right from the start by 4 real-world case studies in political, environmental, social, and sports analytics...essential for learning and great perspective for applying your new skills in your own field. See for yourself how quickly and easily this course can turn you into the data analyst that employers are looking for.

At Course Completion

- Get started with data analysis as quickly and effectively as possible. You'll learn how to use JupyterLab and Jupyter Notebooks to organize and develop your analyses. You'll learn how to use a subset of the Pandas module for data analysis and visualization. And you'll learn how to use a subset of the Seaborn module to create professional data visualizations that can be used for presentations. When you're done with this section, you'll be able to start doing analyses of your own.
- Most analysis is descriptive analysis in which you analyze past data to help you gain new insights. This course presents the critical descriptive analysis skills that you need for success on the job. That includes:
 - Predictive analysis takes data analysis to another level by using statistical models to predict unknown or future values. Although a complete treatment of predictive analysis is far beyond the scope of this class, all analysts should know the basic concepts and skills. That's why this section of this course presents those concepts and gets you started doing your own predictions.
- This section of the course presents 4 case studies that show you how the skills you've been learning can be applied to real-world datasets. Frankly, you can't master on-the-job skills by working with toy datasets, and these case studies help make sure that you will master the professional skills that you need.

Prerequisites

This course is for anyone who wants to become a data analyst, no matter what the field. The only prerequisite is some programming experience, although it doesn't have to be in Python

Contact ISInc for more information at 916.920.1700 or by visiting our website at <http://www.isinc.com>

Student Materials

The student kit includes a workbook and other necessary materials for this class.

Course Outline

Module 1: Introduction to Python for data analysis

- Introduction to data analysis
 - What data analysis is
 - The five phases of data analysis and visualization
 - The IDEs for Python data analysis
- The Python skills that you need for data analysis
 - How to install and import the Python modules for data analysis
 - How to call and chain methods
 - The coding basics for Python data analysis
- How to use JupyterLab as your IDE
 - How to start JupyterLab and work with a Notebook
 - How to edit and run the cells in a Notebook
 - How to use the Tab completion and tooltip features
 - How syntax and runtime errors work
 - How to use Markdown language
 - How to get reference information
- Two more skills for working with JupyterLab
 - How to split the screen between two Notebooks
 - How to use Magic Commands
- Introduction to the case studies
 - The Polling case study
 - The Forest Fires case study
 - The Social Survey case study
 - The Sports Analytics case study

Module 2: The Pandas essentials for data analysis

- Introduction to the Pandas DataFrame
 - The DataFrame structure
 - Two ways to get data into a DataFrame
 - How to save and restore a DataFrame
- How to examine the data
 - How to display the data in a DataFrame
 - How to use the attributes of a DataFrame
 - How to use the `info()`, `nunique()`, and `describe()` methods
- How to access the columns and rows
 - How to access columns
 - How to access rows
 - How to access a subset of rows and columns

- Another way to access a subset of rows and columns
- How to work with the data
 - How to sort the data
 - How to use the statistical methods
 - How to use Python for column arithmetic
 - How to modify the string data in columns
- How to shape the data
 - How to use indexes
 - How to pivot the data
 - How to melt the data
- How to analyze the data
 - How to group the data
 - How to aggregate the data
 - How to plot the data

Module 3: The Pandas essentials for data visualization

- Introduction to data visualization
 - The Python libraries for data visualization
 - Long vs. wide data for data visualization
 - How the Pandas plot() method works by default
 - The three basic parameters for the Pandas plot() method
- How to create 8 types of plot
 - How to create a line plot or an area plot
 - How to create a scatter plot
 - How to create a bar plot
 - How to create a histogram or a density plot
 - How to create a box plot or a pie plot
- How to enhance a plot
 - How to improve the appearance of a plot
 - How to work with subplots
 - How to use chaining to get the plots you want

Module 4: The Seaborn essentials for data visualization

- Introduction to Seaborn
 - The Seaborn methods for plotting
 - The general methods vs. the specific methods
 - How to use the basic Seaborn parameters
 - How to use the Seaborn parameters for working with subplots
- How to enhance and save plots
 - How to set the title, x label, and y label
 - How to set the ticks, x limits, and y limits
 - How to set the background style
 - How to work with subplots
 - How to save a plot

- How to create relational plots
 - How to create a line plot
 - How to create a scatter plot
- How to create categorical plots
 - How to create a bar plot
 - How to create a box plot
- How to create distribution plots
 - How to create a histogram
 - How to create a KDE or ECDF plot
 - How to enhance a distribution plot
- Other techniques for enhancing a plot
 - How to use other Axes methods to enhance a plot
 - How to annotate a plot
 - How to set the color palette
 - How to enhance a plot that has subplots
 - How to customize the titles for subplots
 - How to set the size of a specific plot

Module 5: How to get the data

- How to find the data that you want to analyze
 - Common data sources
 - How to find and select the data that you want
- How to import data into a DataFrame
 - How to import data directly into a DataFrame
 - How to download a file to disk before importing it
 - How to work with a zip file on disk
- How to get database data into a DataFrame
 - How to run queries against a database
 - How to use a SQL query to import data into a DataFrame
- How to work with a Stata file
 - How to get and explore the metadata of a Stata file
 - How to build DataFrames for the metadata and the data
- How to work with a JSON file
 - How to download a JSON file to disk
 - How to open a JSON file in JupyterLab
 - How to drill down into the data
 - How to build a DataFrame for the data

Module 6: How to clean the data

- Introduction to data cleaning
 - A general plan for cleaning the data
 - What the info() method can tell you
 - What the unique values can tell you
 - What the value counts can tell you

- How to simplify the data
 - How to drop rows based on conditions
 - How to drop duplicate rows
 - How to drop columns
 - How to rename columns
- How to find and fix missing values
 - How to find missing values
 - How to drop rows with missing values
 - How to fill missing values
- How to fix data type problems
 - How to find dates and numbers that are imported as objects
 - How to convert date and time strings to the datetime data type
 - How to convert object columns to numeric data type
 - How to work with the category data type
 - How to replace invalid values and convert a column's data type
 - How to fix data problems when you import the data
- How find and fix outliers
 - How to find outliers
 - How to fix outliers

Module 7: How to prepare the data

- How to add and modify columns
 - How to work with datetime columns
 - How to work with string columns
 - How to work with numeric columns
 - How to add a summary column to a DataFrame
- How to apply functions and lambda expression
 - How to apply functions to rows or columns
 - How to apply user-defined functions
 - How lambda expressions work with DataFrames
 - How to apply lambda expressions
- How to work with indexes
 - How to set and remove an index
 - How to unstack indexed data
- How to combine DataFrames
 - How to join DataFrames with an inner join
 - How to join DataFrames with a left or outer join
 - How to merge DataFrames
 - How to concatenate DataFrames
- How to handle the SettingWithCopyWarning
 - What the warning is telling you
 - What to do when the warning is displayed
 - What to watch for when the warning isn't displayed

Module 8: How to analyze the data

- How to create and plot long data
 - How to melt columns to create long data
 - How to plot melted columns
- How to group and aggregate the data
 - How to group and apply a single aggregate method
 - How to work with a DataFrameGroupBy object
 - How to apply multiple aggregate method
- How to create and use pivot tables
 - How to use the pivot() method
 - How to use the pivot_table() method
- How to work with bins
 - How to create bins of equal size
 - How to create bins with equal numbers of values
 - How to plot binned data
- More skills for data analysis
 - How to select the rows with the largest values
 - How to calculate the percent change
 - How to rank rows
 - How to find other methods for analysis

Module 9: How to analyze time-series data

- How to reindex time-series data
 - How to generate time periods
 - How to reindex with datetime indexes
 - How to reindex with a semi-month index
 - How a user-defined function can improve a datetime index
 - How reindexing with an improved index can improve plots
- How to resample time-series data
 - How to use the resample() method
 - How to use the label and closed parameters when you downsample
 - How downsampling can improve plots
- How to work with rolling windows
 - The concept of rolling windows
 - How to create rolling window
 - How to plot rolling window data
- How to work with running total
 - How to create running totals
 - How to plot running totals

Module 10: How to make predictions with a linear regression model

- Introduction to predictive analysis

- Types of predictive models
- Introduction to regression analysis
- How to find correlations between variables
 - The Housing dataset
 - How to identify correlations with a scatter plot
 - How to identify correlations with a grid of scatter plot
 - How to identify correlations with r-values
 - How to identify correlations with a heatmap
- How to use Scikit-learn to work with a linear regression
 - A procedure for creating and using a regression model
 - The function and methods for linear regression models
 - How to create, validate, and use a linear regression model
 - How to plot the predicted data
 - How to plot the residuals
- How to plot regression models with Seaborn
 - The `lplot()` method and some of its parameters
 - How to plot a simple linear regression
 - How to plot a logistic regression
 - How to plot a polynomial regression
 - How to plot a lowess regression
 - How to use the `residplot()` method to plot the residuals

Module 11: How to make predictions with a multiple regression model

- A simple regression model for a Cars dataset
 - The Cars dataset
 - How to create a simple regression model
 - How to plot the residuals of a simple regression
- How to work with a multiple regression model
 - How to create a multiple regression model
 - How to plot the residuals of a multiple regression
- How to work with categorical variables
 - How to identify categorical variables
 - How to review categorical variables
 - How to create dummy variables
 - How to rescale the data and check the correlations
 - How to create a multiple regression that includes dummy variables
- How to improve a multiple regression model
 - How to select the independent variables
 - How to test different combinations of variables
 - How to use Scikit-learn to select the variables
 - How to select the right number of variables

Module 12: The Polling case study

- Get and display the data
- Clean the data
- Prepare the data
- Analyze the data
- More preparation and analysis

Module 13: The Forest Fires case study

- Get and display the data
- Clean the data
- Prepare the data
- Analyze the data
- Use GeoPandas to plot the fires on a map

Module 14: The Social Survey case study

- Introduction to the Social Survey
- The employment data
- The work-life balance data
- How to expand the scope of the analysis
- How to use a hypothesis to guide your analysis

Module 15: The Sports Analytics case study

- Get the data and build the DataFrame
- Clean the data
- Prepare the data
- Plot the summary data
- Plot the shot locations