100 XP

Introduction

2 minutes

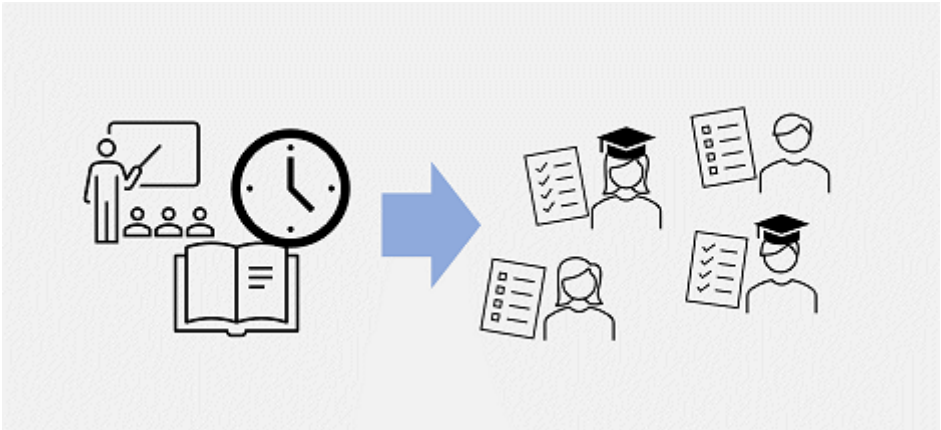
Unsurprisingly, the role of a data scientist primarily involves exploring and analyzing data. The results of an analysis might form the basis of a report or a machine learning model, but it all begins with data, with Python being the most popular programming language for data scientists.

After decades of open-source development, Python provides extensive functionality with powerful statistical and numerical libraries:

- NumPy and Pandas simplify analyzing and manipulating data
- Matplotlib provides attractive data visualizations
- Scikit-learn offers simple and effective predictive data analysis
- TensorFlow and PyTorch supply machine learning and deep learning capabilities

Usually, a data analysis project is designed to establish insights around a particular scenario or to test a hypothesis.

For example, suppose a university professor collects data from their students, including the number of lectures attended, the hours spent studying, and the final grade achieved on the end of term exam. The professor could analyze the data to determine if there is a relationship between the amount of studying a student undertakes and the final grade they achieve. The professor might use the data to test a hypothesis that only students who study for a minimum number of hours can expect to achieve a passing grade.



Prerequisites

- Knowledge of basic mathematics

- Some experience programming in Python

Learning objectives

In this module, you will:

- Common data exploration and analysis tasks.
- How to use Python packages like NumPy, Pandas, and Matplotlib to analyze data

---

Next unit: Explore data with NumPy and Pandas

Continue >