Data Science

Data science is not python, R, SQL, statistics or data visualization. Data science is applying tools to analyze data in order to extract useful information.

There are four jobs within data science:

Job Roles Tools

Data Engineer store & maintain data SQL/Python

Data Analyst visualize & describe data SQL/Bi tools/Excel

Data Scientist Gain insights from data Python/R

Machine learning (ML) Scientist Predict with data Python/R

Data Scientist

Data scientists are responsible for collecting and analyzing complex data sets to identify patterns and insights that inform business strategy. He/she communicates the significance of the data a business holds in a way that others can comprehend.

To achieve this, a data scientist needs to have a working knowledge of the many programming languages and data science methodologies for addressing this challenging task. They must possess a deep understanding of statistical models, machine learning and data mining techniques.

A data scientist role is critical to any organization looking to leverage their data assets for competitive advantage.

Introduction to Data Wrangling with Python

Data science and analytics are taking over the whole world and the job of a data scientist is routinely being called the coolest job of the 21st century. However, for all the emphasis on data, it is the science that makes you – the practitioner – truly valuable. To practice high-quality science with data, you need to make sure it is properly sourced, cleaned, formatted, and pre-processed. This book teaches you the most essential basics of this invaluable component of the data science pipeline: data wrangling.

Data wrangling is the process that ensures that the data is in a format that is clean, accurate, formatted, and ready to be used for data analysis.

After the data scientists identify useful data sources for solving the business problem (for instance, in-house database storage or internet or streaming sensor data), they then proceed to extract, clean, and format the necessary data from those sources.

Data wrangling is generally done at the very first stage of a data science/analytics pipeline.

Generally, the task of data wrangling involves the following steps:

Scraping raw data from multiple sources (including web and database tables)

Imputing, formatting, and transforming – making it ready to for use in the modeling process (such as advanced machine learning)

Handling read/write errors

Detecting outliers

Performing quick visualizations (plotting) and basic statistical analysis to judge the quality of your formatted data.

The process of data wrangling includes first finding the appropriate data that is necessary for the analysis.

This data can be from one or multiple sources, such as tweets, bank transaction statements in a relational database, sensor data, and so on.

This data needs to undergo cleaning. If there is missing data, we will either delete or substitute it, with the help of several techniques.

If there are outliers, we need to first detect them and then handle them appropriately.

If data is from multiple sources, we will have to perform join operations to combine it.

Python for Data Wrangling

There is always a debate on whether to perform the wrangling process using an enterprise tool or by using a programming language and associated frameworks.

There are many commercial, enterprise-level tools for data formatting and pre-processing that do not involve much coding on the part of the user. These examples include the following:

General-purpose data analysis platforms such as Microsoft Excel (with add-ins)

Statistical discovery package such as JMP (from SAS)

Modeling platforms such as Rapid Miner

Analytics platforms from niche players focusing on data wrangling, such as Trifacta, Paxata, and Alteryx

However, programming languages such as Python provide more flexibility, control, and power compared to these off-the-shelf tools. As the volume, velocity, and variety of data undergo rapid changes, it is always a good idea to develop and nurture a significant amount of in-house expertise in data wrangling using fundamental programming frameworks, so that an organization is not beholden to the whims and fancies of any enterprise platform for as basic a task as data wrangling.