

Lesson summary

Module 2 Lesson 1: From Understanding to Preparation



Skills
Network

Congratulations! You have completed this lesson. At this point in the course, you know:

- The Data Understanding stage encompasses all activities related to constructing the data set and answers the question as to whether the data you collected represents the problem to be solved.
- During the Data Understanding stage, scientists might use descriptive statistics, predictive statistics, or both.
- Data scientists commonly apply Hurst, univariates, and other statistics on each variable, such as mean, median, minimum, maximum, standard deviation, pairwise correlation, and histograms.
- Data scientists also use univariates, statistics, and histograms to assess data quality.



Data Understanding and Preparation

Data Understanding and Iterative Assessment

The significant role of data assessment and effective preparation techniques for achieving successful analytical outcomes



IMPORTANCE OF UNDERSTANDING

1

Data understanding ensures the quality along with the representativeness of collected data. It is time-consuming, typically constitutes a significant project duration.



ASSESSMENT AND ITERATION

2

Data understanding assesses the quality and significance of data components. Uses visualizations and descriptive statistics. Iterations refine both problem definition, collection methods



DATA PREPARATION

3

Involves transforming data. Addresses issues like missing values, formatting, duplicates, and feature engineering. It sets the stage for effective model building and analysis.



EFFICIENCY AND QUALITY

4

Automated processes. Prioritize modeling and problem-solving. Feature engineering and text analysis enhance quality and performance. Focus and ensure precise, reliable outcomes.



DATA SCIENCE METHODOLOGY



Skills
Network

- During the Data Preparation stage, data scientists must address missing or invalid values, remove duplicates, and validate that the data is properly formatted.
- Feature engineering, also part of the Data Preparation stage, uses domain knowledge of the data to create features that make the machine learning algorithms work.
- Text analysis during the Data Preparation stage is critical for validating that the proper groupings are set and that the programming is not overlooking hidden data.

Author(s)

[Dr. Pooja](#)
[Patsy R. Kravitz](#)

Changelog

Date	Version	Changed by	Change Description
2023-08-09	0.1	Patsy R. Kravitz	Initial version created
2023-08-13	0.2	Dr. Pooja	Infographic included

© IBM Corporation 2023. All rights reserved.