Glossary

Module 2 Lesson 1: From Understanding to Preparation

Welcome! This alphabetized glossary contains many of the terms you'll find within this lesson. These terms are important for you to recognize when working in the industry, when participating in user groups, and when participating in other certificate programs.

Term	Definition
Automation	Using tools and techniques to streamline data collection and preparation processes.
Data Collection	The phase of gathering and assembling data from various sources.
Data Compilation	The process of organizing and structuring data to create a comprehensive data set.
Data Formatting	The process of standardizing the data to ensure uniformity and ease of analysis.
Data Manipulation	The process of transforming data into a usable format.
Data Preparation	The phase where data is cleaned, transformed, and formatted for further analysis, including feature engineering and text analysis.
Data Preparation	The stage where data is transformed and organized to facilitate effective analysis and modeling.
Data Quality	Assessment of data integrity and completeness, addressing missing, invalid, or misleading values.
Data Quality Assessment	The evaluation of data integrity, accuracy, and completeness.
Data Set	A collection of data used for analysis and modeling.
Data Understanding	The stage in the data science methodology focused on exploring and analyzing the collected data to ensure that the data is representative of the problem to be solved.
Descriptive Statistics	Summary statistics that data scientists use to describe and understand the distribution of variables, such as mean, median, minimum, maximum, and standard deviation.
Feature	A characteristic or attribute within the data that helps in solving the problem.
Feature Engineering	The process of creating new features or variables based on domain knowledge to improve machine learning algorithms' performance.
Feature Extraction	Identifying and selecting relevant features or attributes from the data set.
Interactive Processes	Iterative and continuous refinement of the methodology based on insights and feedback from data analysis.
Missing Values	Values that are absent or unknown in the dataset, requiring careful handling during data preparation.
Model Calibration	Adjusting model parameters to improve accuracy and alignment with the initial design.
Pairwise Correlations	An analysis to determine the relationships and correlations between different variables.
Text Analysis	Steps to analyze and manipulate textual data, extracting meaningful information and patterns.
Text Analysis Groupings	Creating meaningful groupings and categories from textual data for analysis.
Visualization techniques	Methods and tools that data scientists use to create visual representations or graphics that enhance the accessibility and understanding of data patterns, relationships, and insights.

Author(s)

Dr. Pooja
Patsy Kravitz

Changelog

Date	Version	Changed by	Change Description
2023-08-03	0.1	Patsy Kravitz	Initial version created