Advanced Certification Programme in Data Science and Business Analytics with Generative AI

Student Guide for Sync Session

Week 3: Data Cleaning

This guide is your roadmap to making the most of our online session. Packed with essential tips and strategies, it’s designed to keep you engaged, prepared, and ready to dive into a smooth and productive learning journey. Get ready to participate, learn, and thrive!

Session Overview

<table>
<tr><td><b>Session title</b></td><td>Data Cleaning</td></tr>
<tr><td><b>Session duration</b></td><td>3 hours</td></tr>
<tr><td><b>Session type</b></td><td><ul><li><b>Lectures</b>: Conceptual understanding of data cleaning techniques, handling domain-specific challenges, and ensuring data integrity.</li><li><b>Case Studies</b>: Practical application of data cleaning techniques using Python and R.</li></ul></td></tr>
<tr><td><b>Scope</b></td><td>This session introduces the importance of data cleaning in data analysis. It covers:<ul><li>Common data quality challenges.</li><li>Handling missing values and inconsistencies.</li><li>Standardising text and categorical data.</li><li>Preventing data leakage and maintaining data integrity.</li><li>Handling domain-specific data cleaning challenges.</li></ul></td></tr>
<tr><td><b>Learning objectives</b></td><td><table>
<tr><th>Objective</th><th>Core capability</th></tr>
<tr><td>Understand data cleaning principles</td><td>Ability to detect and correct data issues</td></tr>
<tr><td>Apply data quality best practices</td><td>Ensuring accurate and structured datasets</td></tr>
<tr><td>Implement text standardisation techniques</td><td>Improve data consistency for better analysis</td></tr>
<tr><td>Prevent data leakage in machine learning</td><td>Maintain data integrity and avoid overfitting</td></tr>
<tr><td>Handle domain-specific data cleaning challenges</td><td>Apply techniques in finance, healthcare, and e-commerce</td></tr>
</table></td></tr>
<tr><td><b>Software/tools</b></td><td><ul><li>Python (pandas, re, scikit-learn)</li><li>R (dplyr, tidyr, stringr)</li><li>IDE: Jupyter Notebook / Rstudio</li><li>Datasets: Financial, healthcare, e-commerce, and survey data</li><li>Presentation Tool: PowerPoint</li></ul></td></tr>
</table>

Pro Tips for Success

- **Ask Bold Questions:** No question is too small—curiosity is the key to learning!
- **Be Hands-On**: Practice cleaning messy datasets to understand real-world challenges.
- **Collaborate:** Discuss different data quality issues and solutions with peers.

Session Details

Topic	A glimpse	Insight / Actionable
Introduction	Learn why data cleaning is crucial for accurate analysis.	Reflect on how poor data quality affects business insights.
Common Data Quality Challenges	Identify missing values, inconsistencies, and text formatting issues.	Learn techniques to detect and resolve data issues.
Standardising Text and Categorical Data	Understand methods for text cleaning and categorical encoding.	Use Python and R to apply text case standardisation and remove unwanted characters.
Eliminating Irrelevant Data	Learn how to decide which data to keep or remove.	Practice removing unnecessary columns from datasets.
Converting Data Types	Fix issues caused by incorrect data formats.	Convert numerical data stored as text to the correct type.
Preventing Data Leakage	Learn how data leakage affects machine learning models.	Identify and remove features that introduce leakage.
Handling Domain-Specific Data Cleaning Challenges	Explore challenges in financial, healthcare, and e-commerce data.	Address negative financial values, standardise medical codes, and handle currency conversions.
Cleaning Customer Reviews and IoT Data	Remove unwanted characters and incorrect timestamps.	Apply RegEx for text cleaning and filter out invalid timestamps.
Standardising Open-Ended Survey Answers	Learn how to map inconsistent survey responses to predefined categories.	Practice recoding responses for better data analysis.
Best Practices for Data Handling	Summarise key steps to maintain data integrity.	Create a checklist for effective data cleaning.

Post-Session Activities

Reflection challenge	What are the most common data cleaning issues you've encountered?
Explore more	• **Read**: Articles on data quality best practices. • **Watch**: Tutorials on text standardisation and handling missing values. • **Practice**: Apply data cleaning techniques to a raw dataset.
Get inspired	Did you know that data cleaning accounts for up to 80% of the work in data science? Mastering it is a key skill for any data professional!
The journey ahead	Explore advanced topics like data preprocessing for machine learning!