CoGrammar

Welcome to this session

Skills Bootcamp:

Data Manipulation and Cleaning (Theory)

The session will start shortly...

Questions? Drop them in the chat.
We'll have dedicated moderators answering questions.

Safeguarding & Welfare

We are committed to all our students and staff feeling safe and happy; we want to make sure there is always someone you can turn to if you are worried about anything.

If you are feeling upset or unsafe, are worried about a friend, student or family member, or you feel like something isn't right, speak to our safeguarding team:

Ian Wyles
Designated Safeguarding Lead

Simone Botes

Nurhaan Snyman

Rafiq Manan

Ronald Munodawafa

Tevin Pitts

Scan to report a safeguarding concern

or email the Designated Safeguarding Lead:
Ian Wyles
safeguarding@hyperiondev.com

CoGrammar HyperionDev

Skills Bootcamp Full Stack Web Development

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. We will be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

Skills Bootcamp Cloud Web Development

- For all **non-academic questions**, please submit a query: ***www.hyperiondev.com/support***
- **Report a safeguarding incident: *www.hyperiondev.com/safeguardreporting***
- We would love your feedback on lectures: ***Feedback on Lectures.***
- Find all the lecture **content** in your **Lecture Backpack** on GitHub.
- If you are hearing impaired, kindly use your computer's function through Google chrome to enable captions.

Skills Bootcamp Progression Overview

✅ Criterion 1 - Initial Requirements

Specific achievements within the first two weeks of the program.

To meet this criterion, students need to, by no later than **01 December 2024 (C11)** or **22 December 2024 (C12):**

- **Guided Learning Hours** (GLH): Attend a minimum of 7-8 GLH per week (lectures, workshops, or mentor calls) for a total minimum of **15 GLH**.
- **Task Completion:** Successfully complete the **first 4 of the assigned tasks**.

✅ Criterion 2 - Mid-Course Progress

Progress through the successful completion of tasks within the first half of the program.

To meet this criterion, students should, by no later than **12 January 2025 (C11)** or **02 February 2025 (C12):**

- **Guided Learning Hours** (GLH): Complete at least **60 GLH**.
- **Task Completion :** Successfully complete the **first 13 of the assigned tasks**.

CoGrammar

Skills Bootcamp Progression Overview

✅ Criterion 3 – End-Course Progress

Showcasing students' progress nearing the completion of the course.

To meet this criterion, students should:

- **Guided Learning Hours** (GLH)**:** Complete the **total minimum required GLH,** by the **support end date.**
- **Task Completion : Complete all mandatory tasks**, including any necessary resubmissions, by the end of the bootcamp, **09 March 2025 (C11)** or **30 March 2025 (C12).**

✅ Criterion 4 - Employability

Demonstrating progress to find employment.

To meet this criterion, students should:

- **Record an Interview Invite:** Students are required to record proof of invitation to an interview by **30 March 2025 (C11)** or **04 May 2025 (C12)**.
  - **South Holland Students** are required to proof and interview by **17 March 2025.**
- **Record a Final Job Outcome :** Within 12 weeks post-graduation, students are required to record a job outcome.

CoGrammar

Stay Safe Series:

Mastering Online Safety One week at a Time

While the digital world can be a wonderful place to make education and learning accessible to all, it is unfortunately also a space where harmful threats like online radicalization, extremist propaganda, phishing scams, online blackmail and hackers can flourish.

As a component of this BootCamp the *Stay Safe Series* will guide you through essential measures in order to protect yourself & your community from online dangers, whether they target your privacy, personal information or even attempt to manipulate your beliefs.

Don’t Take the Bait:
How to Spot Phishing Scams

- Check the Sender's Email Address
- Look for Generic Greetings
- Be Wary of Urgent Language
- Hover Over Links
- Inspect Attachments Carefully
- Look for Spelling and Grammar Errors
- Verify with the Source
- Use Multi-Factor Authentication
- Stay Informed
- Report Suspicious Emails

Why is data cleaning important in data analysis?

A. To save storage space
B. To ensure accurate and reliable insights
C. To make data visually appealing
D. To remove unused data

What is a common issue when working with datasets?

A. Missing values
B. Overly clear formats
C. Consistently structured data
D. Homogeneous datasets

CoGrammar

Learning Outcomes

- Discuss the importance of data cleaning in the data science pipeline
- Identify common data quality issues and their impact
- Explain key data cleaning techniques
- Explore data manipulation strategies for better analysis.

Lecture Overview

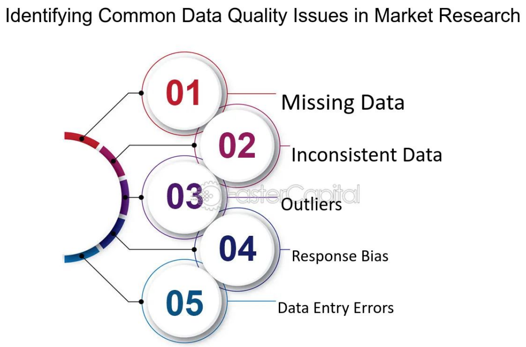
- → Introduction
- → Part 1: Fundamentals of Data Cleaning
- → Break
- → Part 2: Advanced Cleaning and Manipulation
- → Assessment and Q&A

Why Data Cleaning Matters

- 80% of data analysis involves cleaning and preparing data.
- Impact of Dirty Data:
  - Misleading insights.
  - Increased costs and errors.
- Example: Customer IDs missing in a sales dataset.

Common Data Quality Issues

- Missing Data:
  - Null or NaN values in datasets.
- Duplicates:
  - Redundant entries in data.
- Inconsistent Formats:
  - Non-standardized date formats, text inconsistencies.
- Outliers:
  - Extreme values that distort analysis.



Identifying Common Data Quality Issues in Market Research

01 Missing Data

02 Inconsistent Data

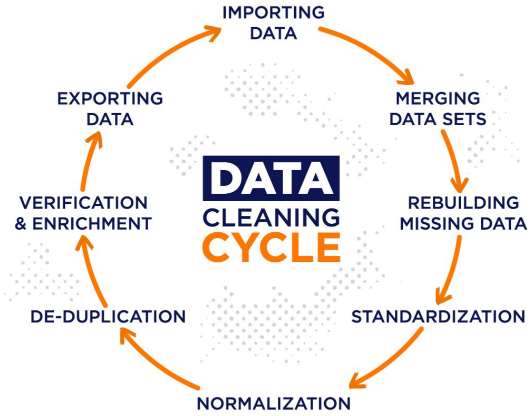
03 Outliers

04 Response Bias

05 Data Entry Errors

Overview of Data Cleaning Process

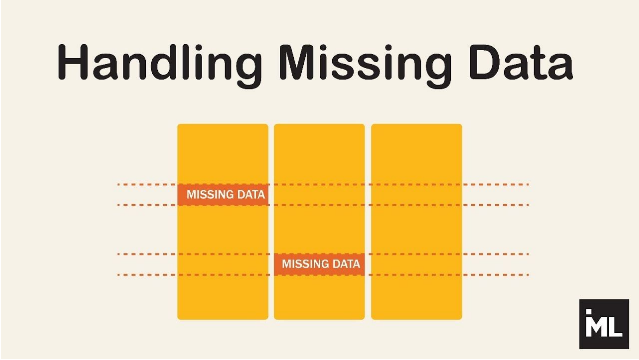
- Identify Issues:
  - Use exploratory data analysis (EDA).
- Handle Missing Data:
  - Impute, drop, or flag missing values.
- Standardize Data:
  - Ensure consistency in formats.
- Remove Duplicates:
  - Retain unique, valid entries.
- Address Outliers:
  - Detect and treat anomalies.



CoGrammar

Handling Missing Data

- Strategies:
  - Deletion: Remove rows/columns with excessive missing values.
  - Imputation: Fill missing data with mean, median, or mode.
  - Flagging: Add an indicator column.
- Example:
  - Missing prices in a product dataset.



Dealing With Duplicates

- Causes:
  - Data entry errors, merging datasets.
- Solution:
  - Identify duplicates using Pandas.DataFrame.duplicated().
  - Drop duplicates while keeping the first or last occurrence.
- Example:
  - Customer orders with redundant IDs.

Standardizing Data Formats

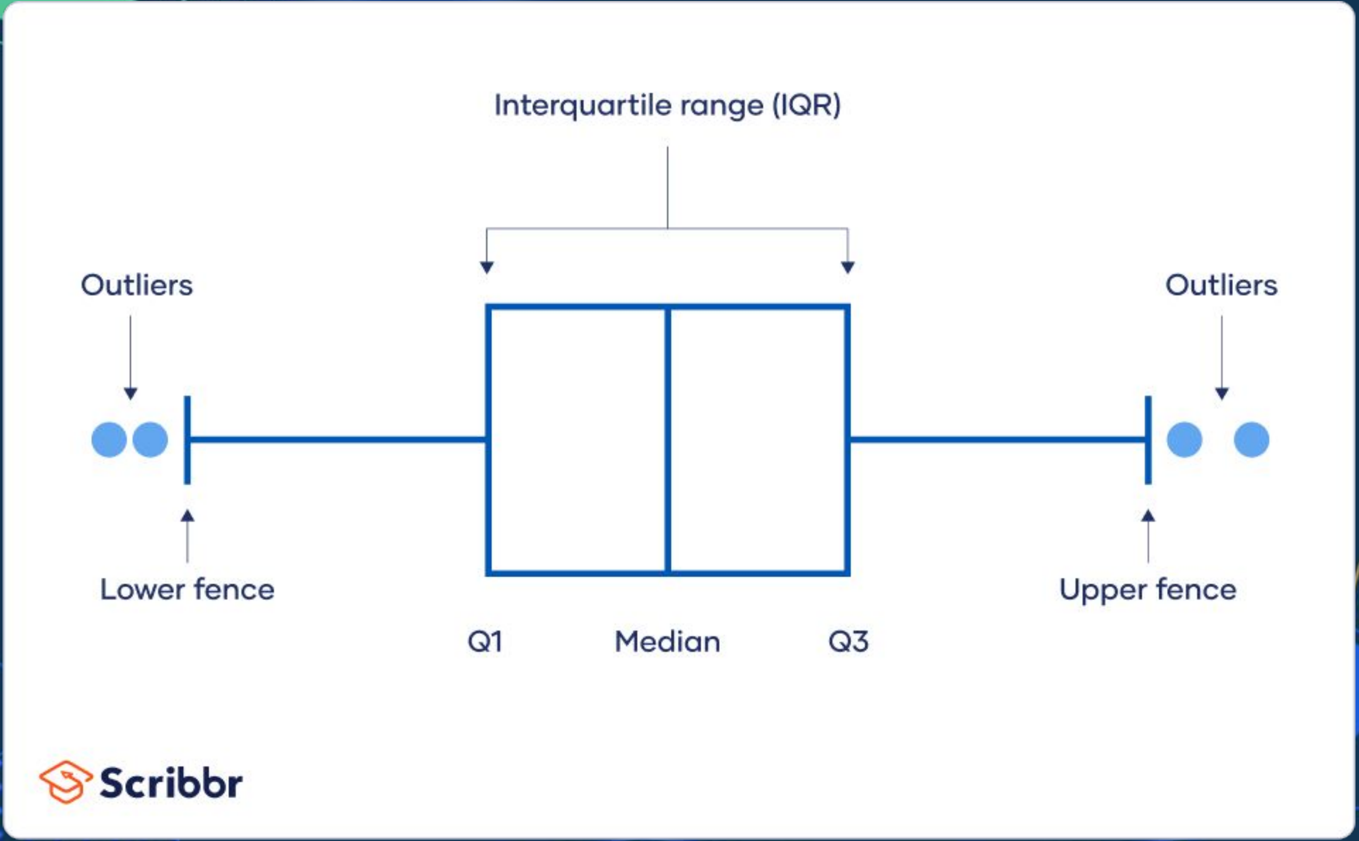
- Focus Areas:
  - Dates: Convert to standard formats (e.g., YYYY-MM-DD).
  - Categorical Variables: Normalize text (e.g., Male/male/M).
  - Numerical Data: Ensure consistent units (e.g., USD vs. EUR).
- Tools:
  - Use Python libraries like Pandas or Excel for transformations.

Outliers

- Outliers are data points that significantly deviate from the rest of the data distribution

CoGrammar

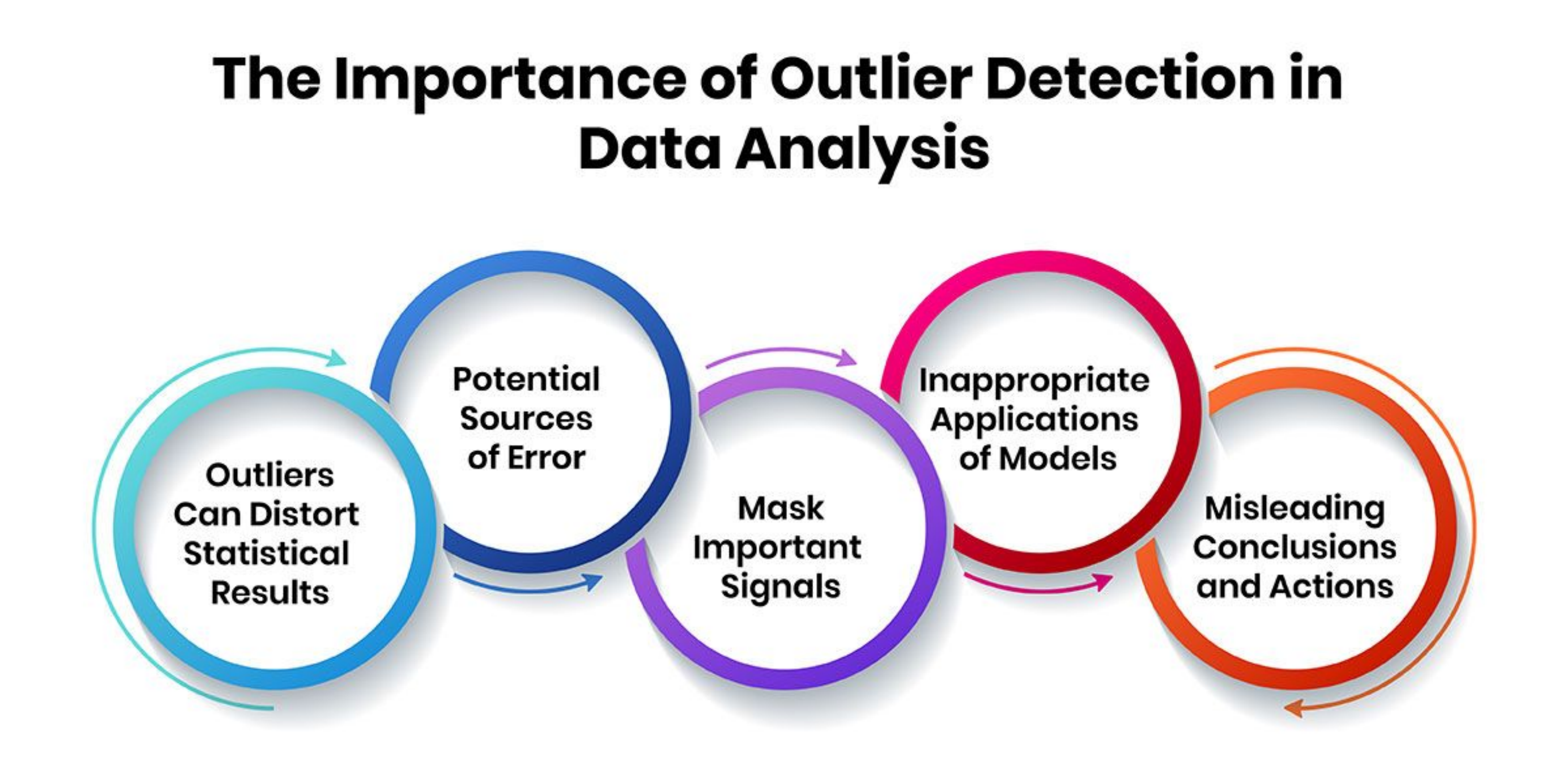
Detecting and Addressing Outliers



CoGrammar

Detecting and Addressing Outliers

- Methods:
  - Visualization: Use boxplots or scatterplots.
  - Statistical Measures: Z-scores, IQR.
  - Solutions: Treat or remove based on context.
- Example:
  - Sales data with an unusually high value.



The Importance of Outlier Detection in Data Analysis

Outliers Can Distort Statistical Results

Potential Sources of Error

Mask Important Signals

Inappropriate Applications of Models

Misleading Conclusions and Actions

Data Manipulation Basics

- Transformations:
  - Normalize or scale data.
- Aggregations:
  - Group and summarize data.
- Combining Datasets:
  - Merge, join, or concatenate tables.

Transforming Data

- Why Transform Data?
  - To prepare it for analysis or machine learning models.
- Examples:
  - Min-max scaling for numerical features.
  - One-hot encoding for categorical variables.

Let's take a break

CoGrammar

Aggregating Data

- Use Cases:
  - Summarizing sales by region.
  - Calculating average customer spend.
- Example:
  - Group data using Pandas’ `groupby` function.

Combining Datasets

- Techniques:
  - Merging: Combine datasets with a common key.
  - Joining: SQL-like operations.
  - Concatenation: Stack datasets vertically or horizontally.
- Example:
  - Merge customer and transaction datasets.

Best Practices for Data Cleaning

- Document Everything:
  - Track changes to data.
- Automate Repetitive Tasks:
  - Use scripts or pipelines.
- Validate Results:
  - Ensure changes improve data quality.
- Maintain Reproducibility:
  - Use version control for data.

Tools for Data Cleaning

- Python Libraries:
  - Pandas: Data manipulation and cleaning.
  - NumPy: Numerical operations.
  - OpenRefine: Cleaning messy data.
- Other Tools:
  - Excel/Google Sheets.
  - R (tidyverse).

Case Study: Cleaning a Dataset

- Dataset:
  - Sales data with missing values, duplicates, and inconsistent formats.
- Step-by-Step:
  - Identify and remove duplicates.
  - Impute missing sales amounts with the median.
  - Standardize date formats.
  - Handle outliers using IQR (Compute the interquartile range).

Real-world Challenges

- Unstructured Data:
  - Text, images, and logs.
- Large Datasets:
  - Memory constraints.
- Messy Data Sources:
  - Legacy systems or scraped data.
- **Tip**:
  - Always tailor cleaning to the context.

Recap Key Points

- Data cleaning ensures accurate and reliable insights.
- Techniques include handling missing data, duplicates, and outliers.
- Data manipulation prepares data for analysis and improves efficiency.

- "Your data is only as good as its quality!"

Which method is best for handling missing data in small datasets?

A. Deleting rows with missing values
B. Imputing values with the dataset mean or median
C. Ignoring the missing data
D. Duplicating rows to fill gaps

What is the best way to detect outliers?

A. Sorting data manually
B. Using boxplots or statistical measures
C. Ignoring extreme values
D. Counting rows with missing values

CoGrammar

Questions and Answers

CoGrammar

Thank you for attending

CoGrammar

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education