Welcome to the CoGrammar

Tutorial: Data Cleaning & Data Preprocessing

The session will start shortly...

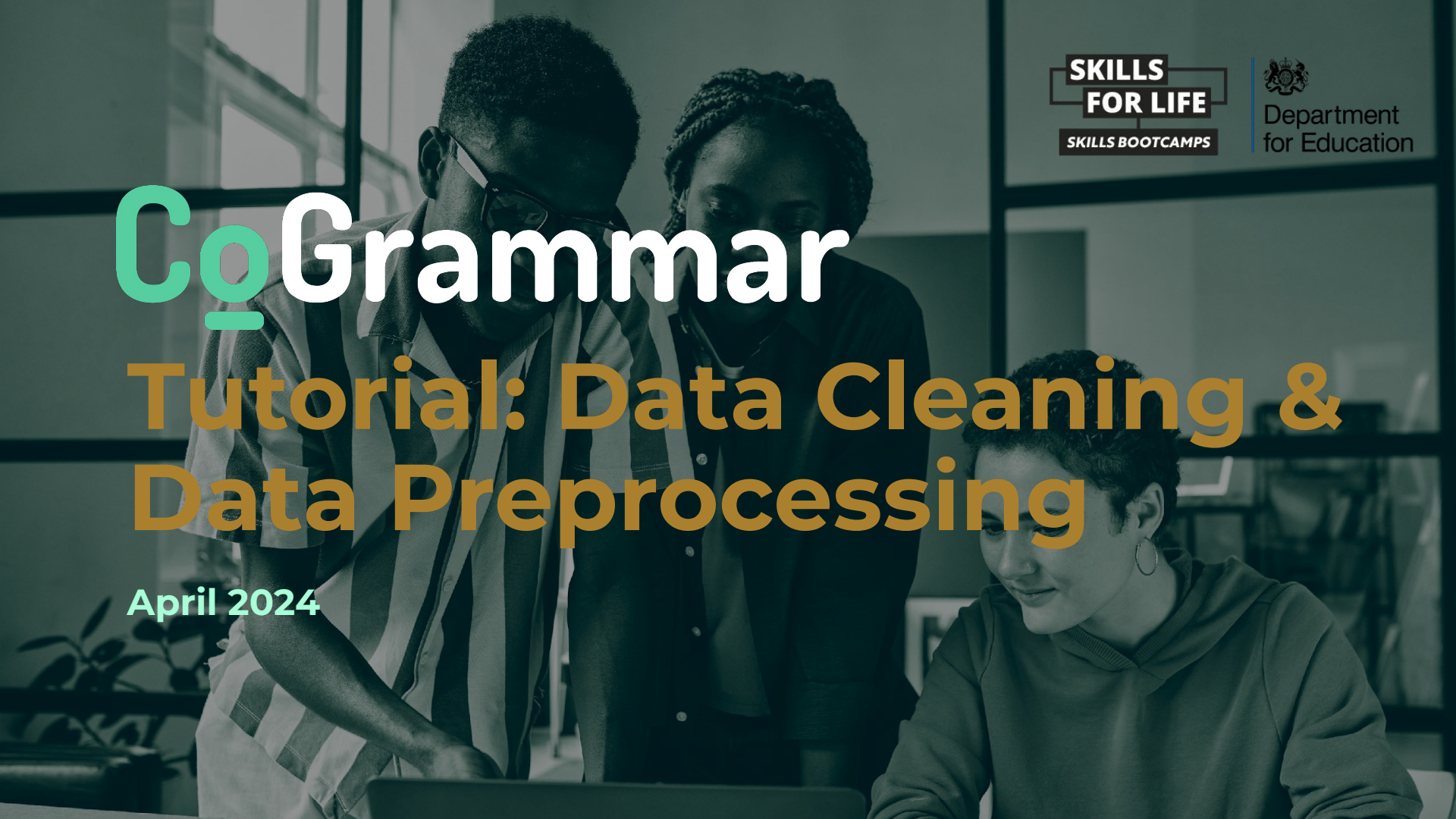
Questions? Drop them in the chat. We'll have dedicated moderators answering questions.

Data Science Session Housekeeping

- 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. Moderators are going to 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**

Data Science Session Housekeeping cont.

- 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**



Tutorial: Data Cleaning & Data Preprocessing

April 2024

Learning objectives

- Explore ways we can tailor our datasets to be better fitted for our goals
- Discuss data cleaning with examples of common errors and inconsistencies

What is the purpose of data normalisation?

A. To convert data into a standard format
B. To handle missing data
C. To remove outliers
D. To perform feature selection

What is the purpose of data normalisation?

A. To convert data into a standard format
B. To handle missing data
C. To remove outliers
D. To perform feature selection

Z-score normalisation transforms features to have a mean of 0 and unit variance.

A. True

B. False

CoGrammar

Z-score normalisation transforms features to have a mean of 0 and unit variance.

A. True

B. False

CoGrammar

Min-max scaling is the optimal feature scaling technique to use when data follows a Gaussian distribution.

A. True

B. False

Min-max scaling is the optimal feature scaling technique to implement when data follows a Gaussian distribution.

A. True

B. False

Which string method is commonly used for removing leading and trailing whitespace from a string?

A. split()

B. replace()

C. strip()

D. upper()

Which string method is commonly used for removing leading and trailing whitespace from a string?

A. split()

B. replace()

C. strip()

D. upper()

Which of the following is not a data formatting issue?

A. Inconsistent capitalisation across entries

B. Dates represented in different formats

C. Missing values in numerical columns

D. Irregular data entry conventions

Which of the following is not a data formatting issue?

A. Inconsistent capitalisation across entries

B. Dates represented in different formats

C. Missing values in numerical columns

D. Irregular data entry conventions

Questions and Answers

CoGrammar

Thank you for attending

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar