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.

CoGrammar

Data Science Session Housekeeping

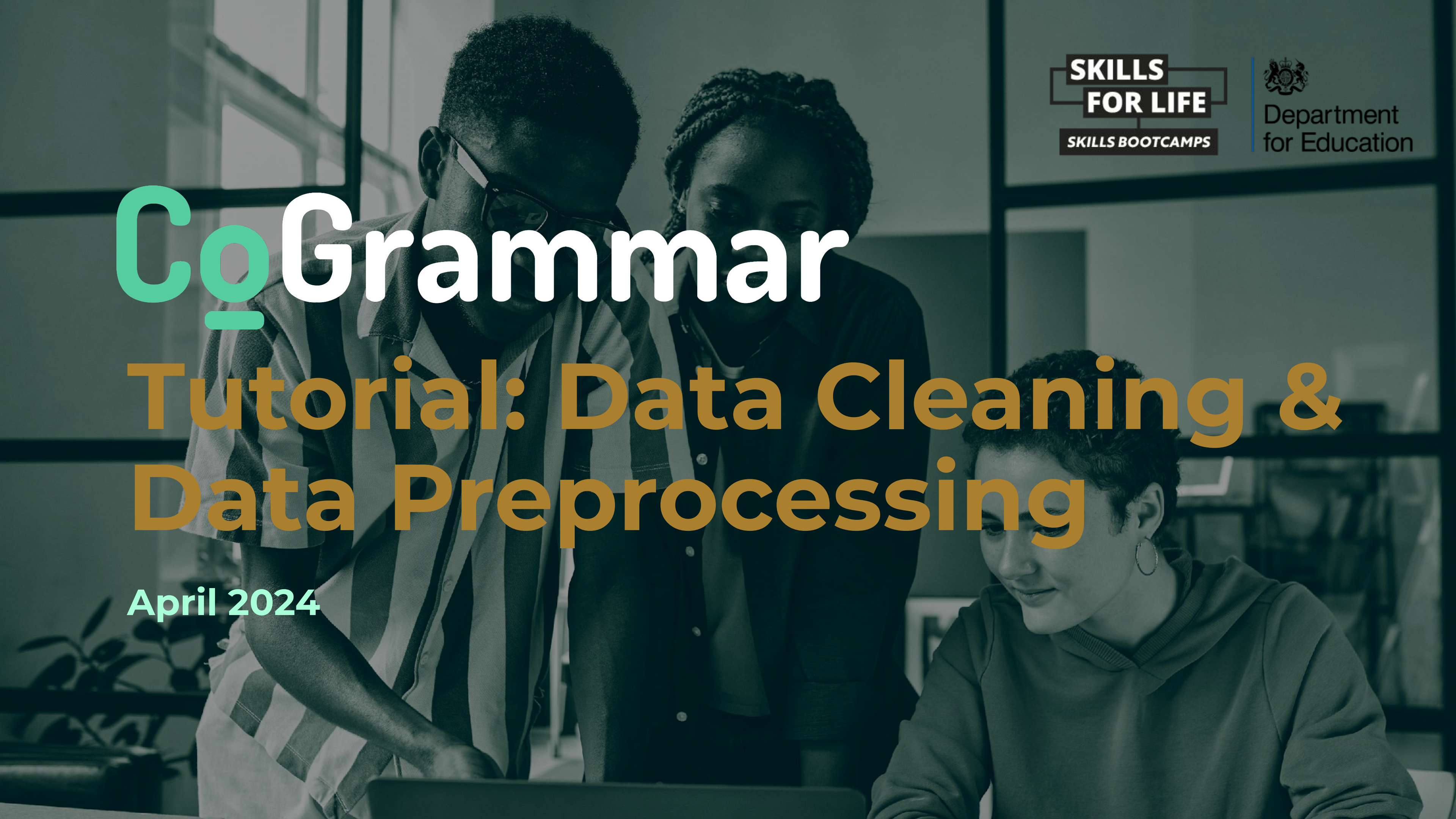
- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at 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



SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

CoGrammar

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

Which of the following is NOT a common data quality issue addressed in data cleaning?

A. Missing values

B. Duplicates

C. Inconsistent formatting

D. Imbalanced classes

CoGrammar

Which of the following is NOT a common data quality issue addressed in data cleaning?

A. Missing values

B. Duplicates

C. Inconsistent formatting

D. Imbalanced classes

CoGrammar

Which technique is suitable for handling missing data only if the amount is minimal and missing completely at random (MCAR)?

A. Mean imputation

B. Deletion

C. K-Nearest Neighbors imputation

D. Multiple Imputation by Chained Equations

Which technique is suitable for handling missing data only if the amount is minimal and missing completely at random (MCAR)?

A. Mean imputation

B. Deletion

C. K-Nearest Neighbors imputation

D. Multiple Imputation by Chained Equations

In Pandas, which function can be used to identify duplicate records in a dataset?

A. find_duplicates()

B. duplicated()

C. is_duplicate()

D. has_duplicates()

In Pandas, which function can be used to identify duplicate records in a dataset?

A. find_duplicates()

B. **duplicated()**

C. is_duplicate()

D. has_duplicates()

Which of the following is a technique for standardizing inconsistent text case?

A. astype()

B. to_datetime()

C. str.upper() or str.lower()

D. strip()

CoGrammar

Which of the following is a technique for standardizing inconsistent text case?

A. astype()

B. to_datetime()

C. str.upper() or str.lower()

D. strip()

Which strategy replaces outlier values with the nearest non-outlier values?

A. Removal
B. Transformation
C. Winsorization
D. Standardization

Which strategy replaces outlier values with the nearest non-outlier values?

A. Removal

B. Transformation

C. Winsorization

D. Standardization

What is the purpose of feature scaling?

A. To convert categorical variables into numerical representations

B. To create new features from existing data

C. To ensure fair comparison and contribution of features in machine learning

D. To handle imbalanced class distributions

What is the purpose of feature scaling?

A. To convert categorical variables into numerical representations

B. To create new features from existing data

C. To ensure fair comparison and contribution of features in machine learning

D. To handle imbalanced class distributions

What is the main difference between nominal and ordinal variables?

A. Nominal variables have categories with an inherent order, while ordinal variables do not

B. Ordinal variables have categories with a meaningful order, while nominal variables do not

C. Nominal and ordinal variables are the same

D. Nominal variables are always encoded using one-hot encoding, while ordinal variables use label encoding

What is the main difference between nominal and ordinal variables?

A. Nominal variables have categories with an inherent order, while ordinal variables do not

B. Ordinal variables have categories with a meaningful order, while nominal variables do not

C. Nominal and ordinal variables are the same

D. Nominal variables are always encoded using one-hot encoding, while ordinal variables use label encoding

Which encoding technique is useful when the frequency of categories is informative?

A. One-hot encoding

B. Label encoding

C. Frequency-based encoding

D. Target encoding

Which encoding technique is useful when the frequency of categories is informative?

A. One-hot encoding

B. Label encoding

C. Frequency-based encoding

D. Target encoding

CoGrammar

Which technique is used to generate synthetic minority instances in imbalanced data handling?

A. Random undersampling

B. Random oversampling

C. SMOTE

D. Tomek Links

CoGrammar

Which technique is used to generate synthetic minority instances in imbalanced data handling?

A. Random undersampling

B. Random oversampling

C. SMOTE

D. Tomek Links

Questions and Answers

Thank you for attending

SKILLS FOR LIFE

SKILLS BOOTCAMPS

Department for Education

CoGrammar