Exploratory Data Analysis

(EDA)

Exploring the dataset

Profiles of customers for an online retail company.

- How many customers?

- How many different products?

- What country spends how much?

- Company's profit last year?

- What period does the data span?

CAMBRIDGE SPARK

Preprocessing data: the important steps

- Categorical features: how to deal with non numerical values?

	Country
Name	
Nick	UK
Laura	IT
Massimo	IT
Paul	FR

CAMBRIDGE SPARK

Preprocessing data: One Hot Encoding

Transforming categorical features into a set of binary features

Name	Country
Nick	UK
Laura	IT
Massimo	IT
Paul	FR

→

Name	Country_FR	Country_IT	Country_UK
Nick	0	0	1
Laura	0	1	0
Massimo	0	1	0
Paul	1	0	0

CAMBRIDGE SPARK

Preprocessing data: the important steps

- Categorical features: how to deal with non numerical values?

- Imputation: what to do with NaN values?

Preprocessing data: the important steps

- Categorical features: how to deal with non numerical values?

- Imputation: what to do with NaN values?

- Outliers: what to do with rows with extreme values?

CAMBRIDGE SPARK

Preprocessing data: the important steps

- Categorical features: how to deal with non numerical values?

- Imputation: what to do with NaN values?

- Outliers: what to do with rows with extreme values?

- Scaling: weight features equally

CAMBRIDGE SPARK

Hands-on session

eda.ipynb