CoGrammar

Data Preprocessing & Exploratory Data Analysis

SKILLS FOR LIFE
SKILLS BOOTCAMPS

Department for Education

Data Science Lecture Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(FBV: Mutual Respect.)**
- 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 Open Classes. You can submit these questions here: Open Class Questions

Data Science Lecture 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**

Lecture Objectives

- Recapping data cleaning & preprocessing.
- Introducing and explaining normalisation and standardisation.
- Introduce the exploratory data analysis.

Recap : Data Analysis

★ When trying to understand a dataset, we need to think analytically.

★ However, it might not be enough and we would need the right tools to analyse our dataset.

★ With the main goal being that we gain an in-depth understanding of our dataset.

Step One : Clean our Data

★ Data that has inconsistencies / missing values, introduces noise, which affects our analysis.

★ There are techniques we can implement to clean our data:
  - Dropping columns
  - Filling and replacing
  - Indexing
  - Creating sub-datasets

Dropping Columns

first_name	last_name	student_id	gender	math_score	english_scor	science_scor	Mother	Father	Address
Theresa	Imore	1	F	89	78	50	Jane Claremint	Michael Scofield	90 East Buckingham
Brook	Gratrex	2	M	67	56	65	Joesephine Brow	Lincon Burrows	60 Fairway Ave.
Jerrold	Isenor	3	M	98	67	42	Claire Mint	Ballack Benet	749 West Kingston Street
Jo	Pretsel	4	M	56	72	87	Alexis Mahone	Duke Harry	Lawrence Township, NJ

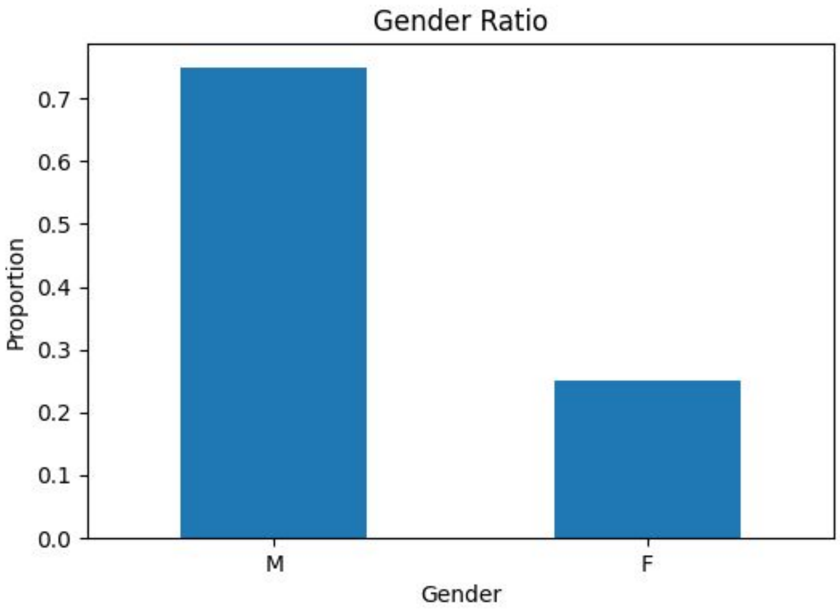
★ **Dropping columns depends on our goal of our analysis. For instance, we want to analyse the proportion of Male / Female, location data. We wouldn't need grades in this case.**

```
data.drop(['math_score', 'english_score', 'science_score'],
          axis=1, inplace=True)
```

- Here is the result of us using the dataset after dropping the redundant columns.

- As you can see, we only drop data that is not useful.

  Typically we try not to drop data, so this method comes in when it becomes very obvious the data isn't needed.



CoGrammar

Filling and Replacing

- ★ Datasets frequently contain missing values and exhibit inconsistencies due to varied data entry practices. We want to ensure data consistency and keep entries uniform.
- ★ For example, an individual's nationality could be entered as “UK”, while others are entered as “United Kingdom”. In this case we opt to stick with “United Kingdom” in the data. If they are not from the UK, we label it “Other”

```
data.replace('UK', 'United Kingdom', inplace=True)
data.fillna('Other', inplace=True)
```

Indexing & Creating sub-datasets

- ★ Often, we want to be able to find out if a specific rule works generally. If we can make a prediction on seen data, how well would it hold up to unseen data?
  - This concept will be looked at further in Machine Learning.
- ★ It would be necessary to split up our dataset:

```
train_data = data.loc[1:80]
test_data = data.loc[81:100]
```

- ★ Common splits are 80:20 for train:test in industry.

Step Two : Missing Data

- The method .fillna() can be dangerous tool. If done incorrectly, it is possible to introduce bias to our dataset.

- How do we use this responsibly? That would depend on the type of missingness.
  - Missing Completely at Random (MCAR)
  - Missing at Random (MAR)
  - Missing Not at Random (MNAR)

Missing Completely at Random (MCAR)

- Missing data is not in any way affected by other variables in the dataset.
  - E.g. A test paper was lost, or a blood sample is missing.
  - There is no reasonable way to guess the missing value.

Missing at Random (MAR)

- ★ Missing data is in some way related to other variables in the dataset.
- ★ Arises as a result of implicit biases in the data itself.
    - ○ E.g. Consider a survey on depression. Statistically, males are less likely to fill in a survey on the severity of their depression.
    - ○ As a result, missing data can be guessed from the proportion of males in the survey.

Missing Not at Random (MNAR)

- ★ Similar to MAR, it arises as a result of implicit biases.
- ★ However, these biases haven't been measured on the data itself.
  - E.g. Fewer reported COVID cases once restrictions have been lifted.
  - It does not mean COVID has disappeared, simply means that less people are getting tested.
- ★ Different to MCAR, as it is not random event that caused the data to go missing.

Let’s Breathe

Let’s take a small break before moving on to the next topic.

CoGrammar

Standardisation & Normalisation

★ Useful when studying relationships between variables in a dataset.

★ Units of measurement can sometimes affect your observations.

★ This also affects Machine Learning applications.

★ Also known as Feature Scaling.
  - Feature scaling is the process of restricting the values in a particular feature to ensure that all features lie on the same scale.

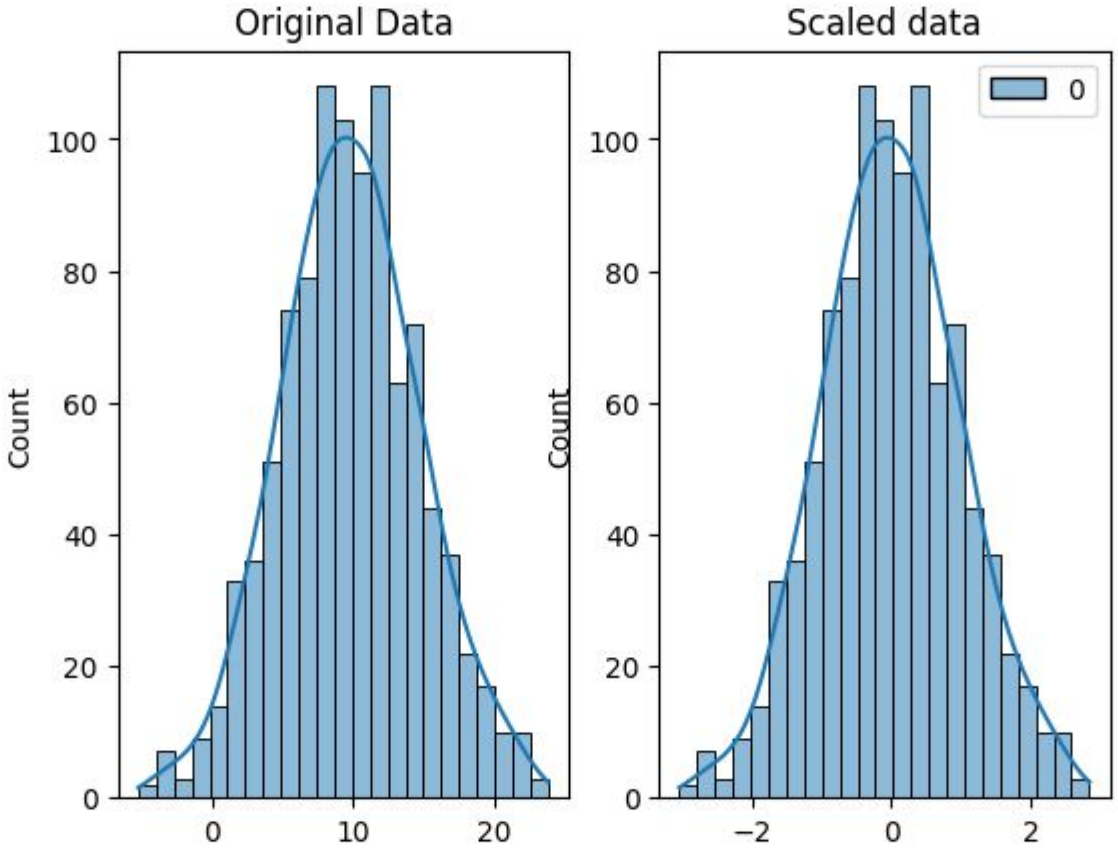
What's the Difference?

It's common to confuse standardization with normalization because these terms are often used interchangeably, and their techniques are somewhat similar. Both processes transform the values of numeric variables to improve the interpretability or comparability of the data.

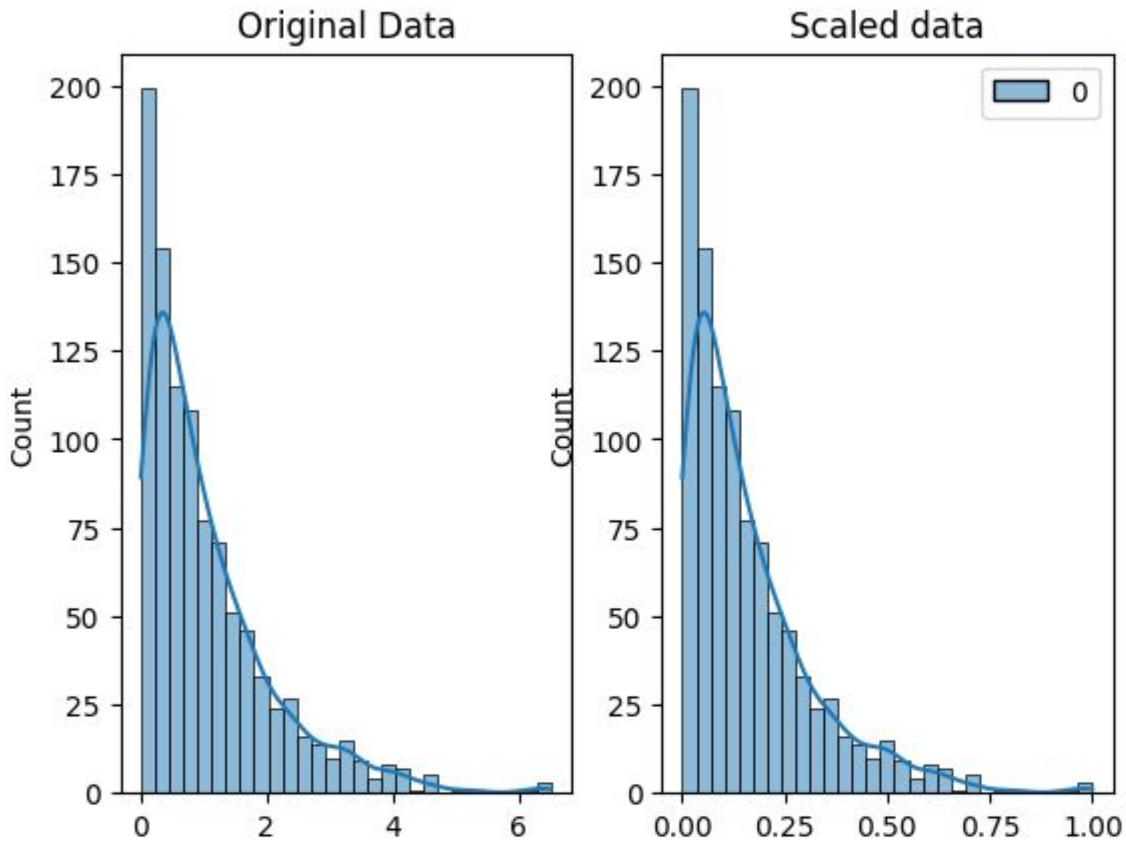
Standardization (Z-score normalization) adjusts the data so that it has a mean (average) of 0 and is distributed around the mean. This scales the data while maintaining the original distribution's shape and is particularly useful when the data follows a Gaussian distribution.

Normalization, on the other hand, refers to a variety of scaling techniques, with min-max scaling being the most common. In min-max scaling, the data is scaled to a fixed range, usually 0 to 1. This means that the minimum value of the data is transformed to 0, and the maximum value is transformed to 1. It is useful when you need to bound your data within certain limits.

By employing these techniques, we ensure that the scale of the data does not distort the results of analysis, particularly when using algorithms that are sensitive to the magnitude of values.



Original Data
Scaled data
0
Count
Count
100
80
60
40
20
0
0
10
20
−2
0
2



Original Data
Count
200
175
150
125
100
75
50
25
0
0
2
4
6
Scaled data
0
Count
200
175
150
125
100
75
50
25
0
0.00
0.25
0.50
0.75
1.00

It is important to note that standardization and normalization only change the scale, not the shape, of the distribution because algorithms that rely on the scale of the data will behave differently, while the inherent data structure remains unchanged.

Exploratory Data Analysis

★ EDA is a crucial step in the data science process where we visually and quantitatively explore data to understand its main characteristics.

★ This step informs us about the underlying structure, detects outliers and anomalies, and tests assumptions before applying further statistical or machine learning models.

★ Through EDA, we uncover patterns, spot anomalies, frame hypotheses, and ensure the data's appropriateness for further analysis.

CoGrammar

How to Explore your Dataset

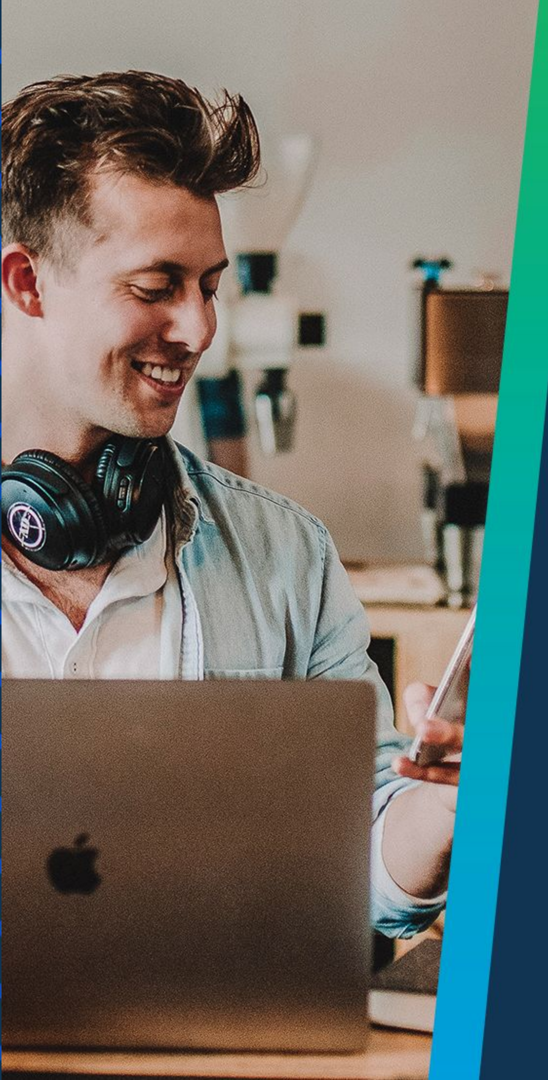
- ★ There is no single prescribed method for EDA; instead, use a variety of statistical and visualization techniques that best uncover the insights within your dataset.

- ★ Key questions to guide your EDA include:
  - What type of data do I have (e.g., numerical, categorical)?
  - Where are the missing values, and how should they be addressed?
  - Where are the outliers, and what might they indicate?
  - How can I modify features to better capture the essence of the data?

Q & A SECTION

Please use this time to ask any questions relating to the topic, should you have any.

CoGrammar



Thank you for joining us

1. Take regular breaks
2. Stay hydrated
3. Avoid prolonged screen time
4. Practice good posture
5. Get regular exercise

"With great power comes great responsibility"