PROJECT TOOLS & DATA WRANGLING

UCL DATA SCIENCE SOCIETY
PROJECT LABS 2026

SESSION 3 | 10 Feb 2026

PROJECT PIPELINE

Where are we at

Step 1:
Problem Definition
(5%)

Step 2:
Collecting data
(10%)

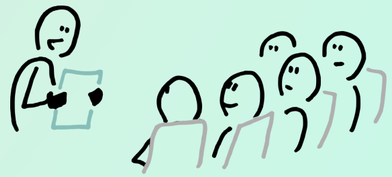
Step 3:
Data Preprocessing
(15%)

Step 4:
EDA
(20%)

Step 5:
Predictive Modelling
(20%)

Step 6:
Visualization &
Interpretation
(20%)

Step 7:
Presentation
(10%)



PROJECT TOOLS

Purpose

Create a Reproducible Project through **Synchronisation, Traceability** and **Accountability**

Key Activities

- Version Control
- Project Management/Task Tracking
- Documentation

GIT/GITHUB - VERSION CONTROL

Purpose

Have one single **True Source** of your project

Key Activities

- Staging and Committing – Saving changes
- Pulling/Pushing – Syncing local work with a remote server
- Branching and Merging– Creating seperate version of projects to experiment

Outcome

Reproducibility, Conflict Resolution and **Professional Transparency**

GitHub

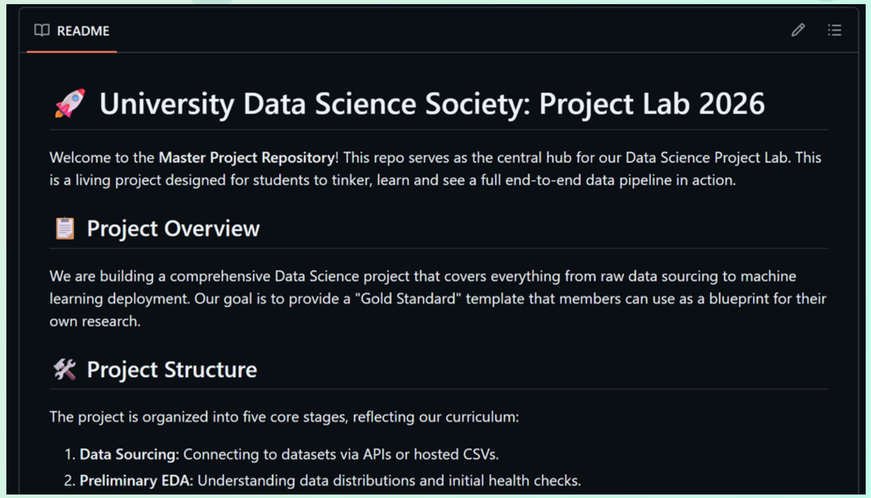
GITHUB

OPTIONAL

SET UP TEAM GITHUB

CREATE A GITHUB REPOSITORY FOR YOUR TEAM

- Add a code file that you can all work on
- Store your data in one place
- Great for visibility with Recruiters



DATA WRANGLING

FROM DATA SOURCING TO ML READY DATA

Definition

The Bridge between Data **Aquisition** and Data **Modelling** → the action taking to resolve issues found in EDA

Benefits

- Handling Issues with Datasets fairly
- Standardised Data Treatment
- Improved Model Accuracy through better quality data

WHAT WE DISCOVERED IN EDA

1. **Data Shape and Structure**
2. **Inspecting Variables and their relationships**
3. **Identifiying Red Flags**

POSSIBLE RED FLAGS

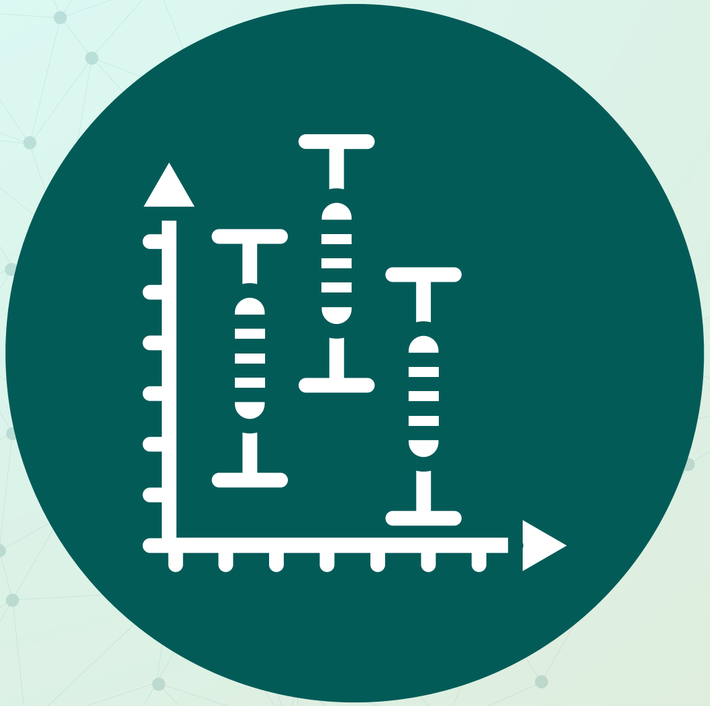
1. OUTLIERS
2. NULL VALUES
3. NAMING CONVENTIONS

```
--- Missing Values ---
Country                0
Temperature           10
Weather_Condition     10
dtype: int64

--- Numerical Summary ---
       Temperature
count    90.000000
mean     26.552222
std     158.285726
min    -999.000000
25%      15.750000
50%      22.000000
75%      28.725000
max     999.900000

--- Unique Countries ---
['Mexico' 'USA' 'Canadâ' 'Canada' 'mexico' ' CANADA ' ' USA ' 'MEXICO '
 'uSA']
```

GAME TIME
FIND THE ISSUES

Order_ID	Date	Customer_Name	Age	Country	Total_Spent	Order_Status
1001	1/25/2023	Alice Johnson	28	USA	$150.00	Shipped
1002	14/02/2023	bob smith	-5	United States	$45.00	Shipped
1003	3/1/2023	Charlie P.	45	CAN	NaN	Pending
1004	1/25/2023	Alice Johnson	28	USA	$150.00	Shipped
1005	5/12/2023	Diana Prince	32	UK	$21,000.00	Cancelled
1006	6/15/2023	Edward Nygma	145	Mexico	$12.50	Delivered
1007	7/20/2023	Fiona Hill	30	u.s.a.	$88.00	Unknown
1008	8/22/2023	George Miller	35	Canada	-$10.00	Delivered

1 HANDLING OUTLIERS



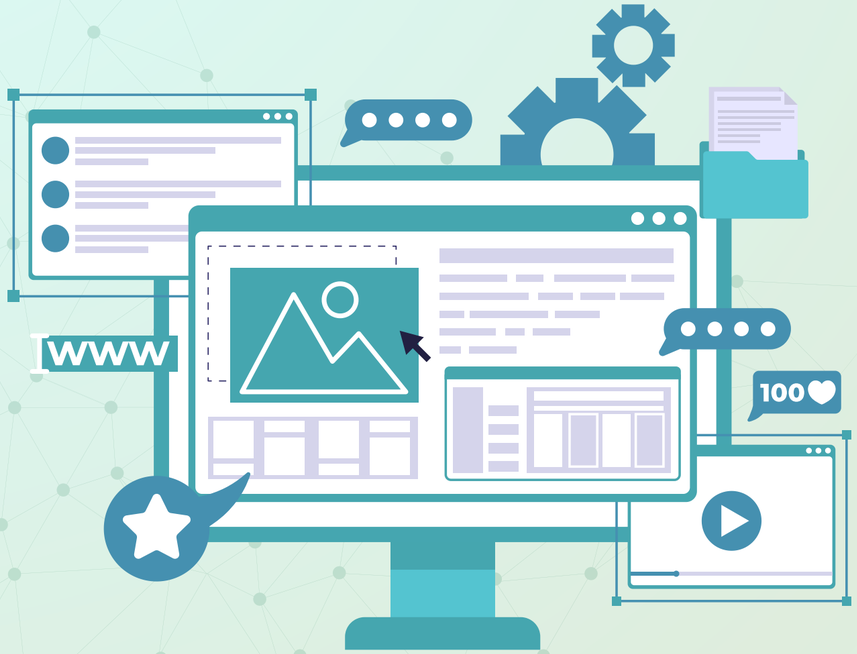
How do we find them?

- Domain Knowledge (e.g. Age 120 is rare)
- Statistical Distribution

How do we deal with them?

- Removing them
- Capping it at the 99th percentile
- Mathematical Transformation
- Imputation – treat it like a missing value

2 HANDLING NULL VALUES



Detection & Diagnosis

- Scattered Missingness or systematic pattern?
- Critical vs Non Critical

How do we deal with them?

- Removal
- Basic Imputation
- Model Based Imputation (knn, etc.)

3 NAMING CONVENTION ISSUES

Detection & Diagnosis

- Inconsistent casing
- Illegal characters
- Ambiguity

How we deal with them?

- Standardisation
- Descriptive Renaming
- Adding Units
- Categorical Value Alignment

PRACTICAL PORTION: DATA WRANGLING

3. Data Wrangling & Quality Control

In this session, we will explore the essential steps of data cleaning. High-quality analysis depends on clean data, but real-world datasets are often "noisy."

We are working with a **synthetic weather dataset** specifically designed to showcase common data integrity issues. Your goal is to identify and resolve these problems using `pandas`.

Key Issues to Address:

- **Outliers**: Physically impossible temperature readings that skew statistical averages.
- **Null Values**: Missing data points that can break analytical functions.
- **Naming Inconsistencies**: Variations in spelling, casing, and spacing for categorical data (Country and Weather Condition).

[Link to the Google Colab – Data Wrangling]

Google colab

ML BASICS

Purpose

Use existing information for prediction or classification tasks

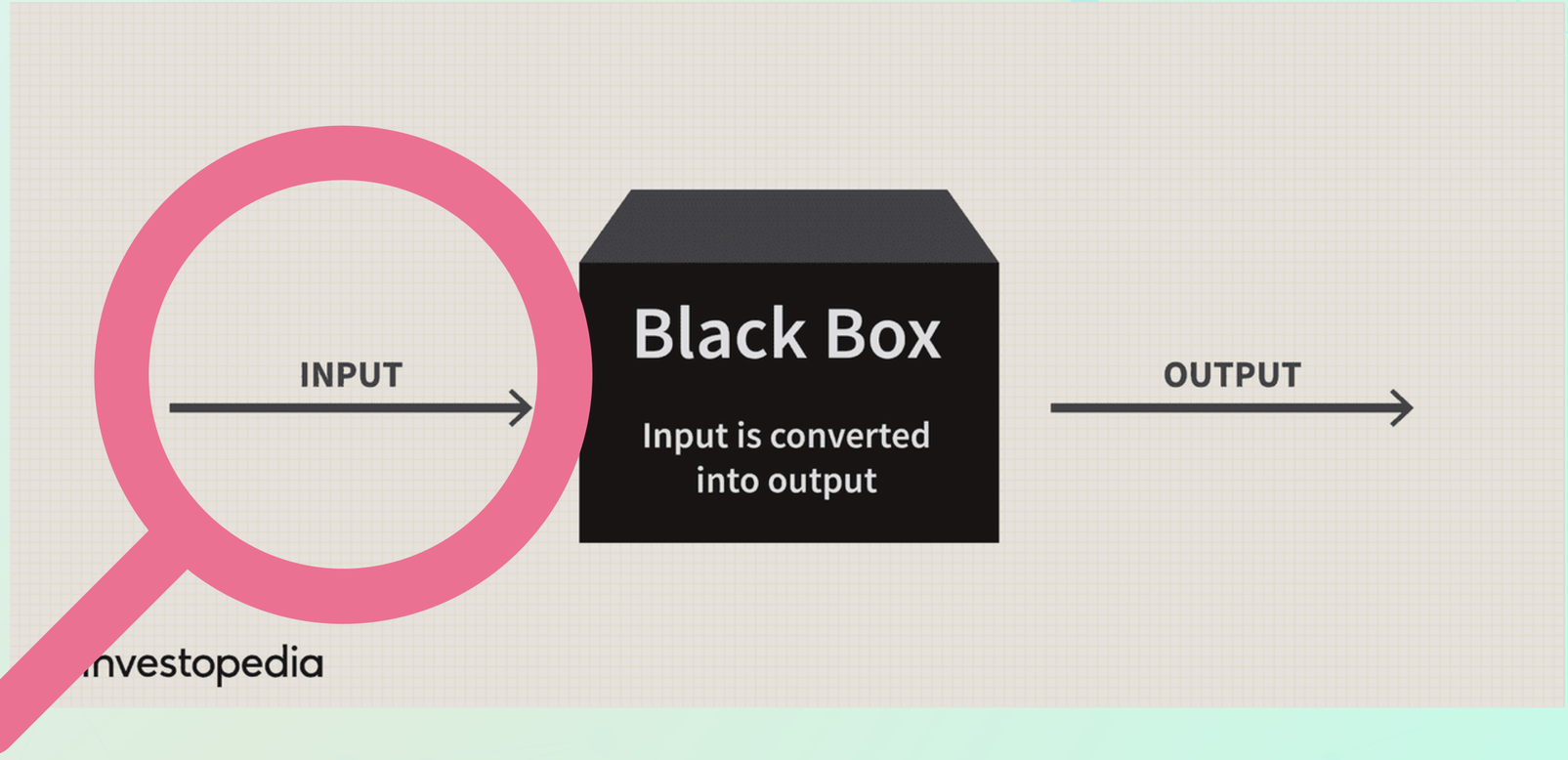
Main Process

- Data Wrangling & Preprocessing: Prepare the dataset for modelling
- Feature Selection: Select the appropriate features for modelling
- Model Selection: Choose the appropriate algorithm
- Modelling: Use existing information (Features) to make a prediction
- Model Evaluation: Evaluate performance using suitable metrics

Importance of Preprocessing

- Improves model performance and reliability
- Ensures data is usable by ML algorithms
- Allow models to learn meaningful patterns

ML BASICS



We are going to focus on the input today

DATA PREPROCESSING

Purpose

Work towards **Numerical Alignment, Information Transformation** and **Efficiency**

Key Activities

- Feature Scaling
- Feature Engineering and Selection
- Categorical Encoding

Outcome

Transforming the dataset to have higher predictive power by being more compatibile for Machine Learning

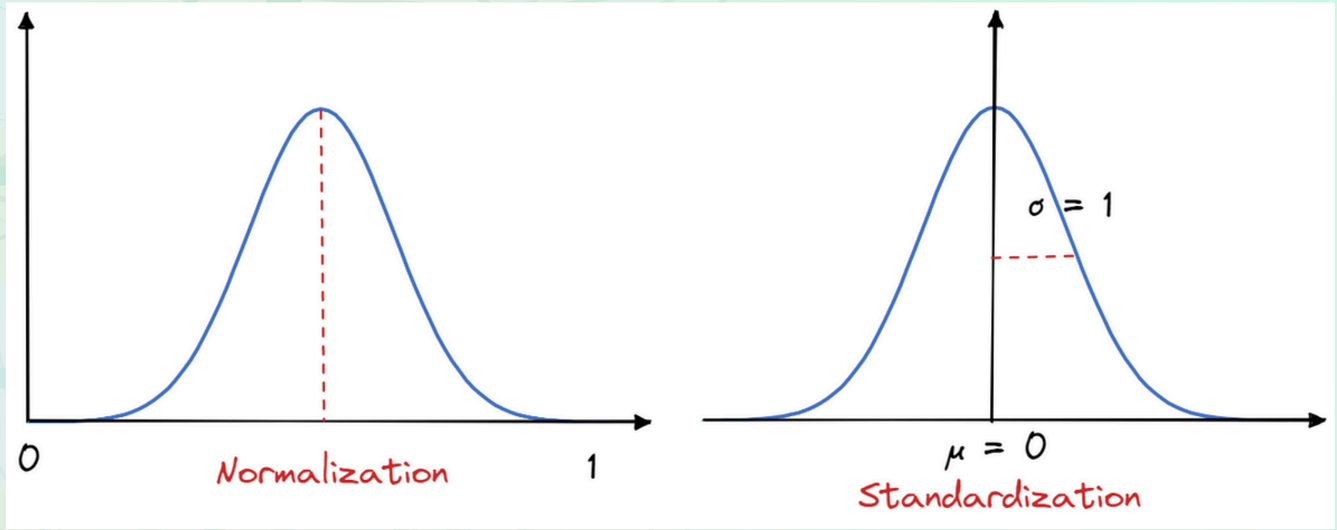
NORMALISATION VS STANDARDISATION

NORMALISATION

→ MIN-MAX SCALING

STANDARDISATION

→ Z-SCORE SCALING



WHEN TO USE WHICH?

NORM → NEED VALUES IN A FIXED RANGE, NOT NORMALLY DISTRIBUTED DATA

STAND → NORMAL DISTRIBUTION AROUND SOME CENTER, GOOD FOR OUTLIERS

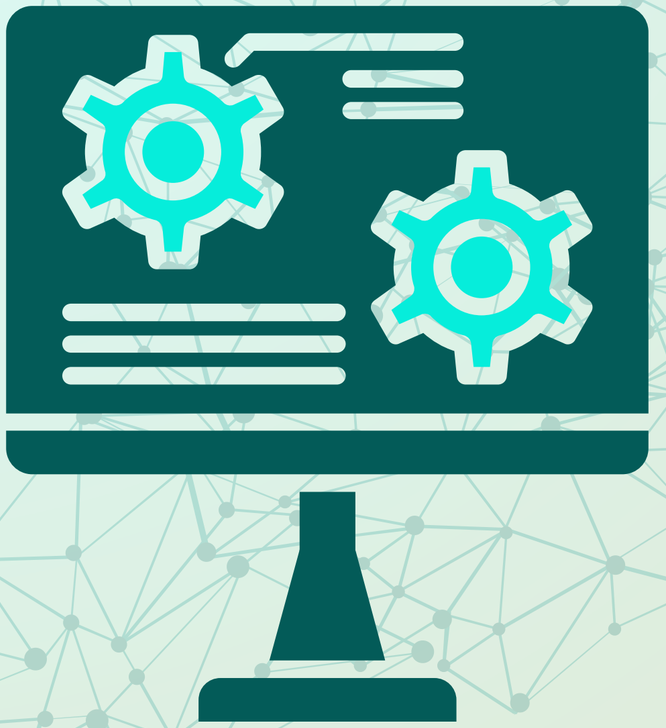
CATEGORICAL ENCODING

City	Weather Conditions
London	Snowy
Dublin	Rainy
Paris	Snowy

Same Information, just a different way to represent it

City	Is_Snowy	Is_Rainy
London	True	False
Dublin	False	True
Paris	True	False

FEATURE ENGINEERING



PURPOSE

- Capture meaningful information in a new feature

KEY ACTIVITIES

- Decomposition (ie. breaking timestamp down)
- Aggregation (combining features)
- Interaction Features

EXAMPLES

- Combining date and time into timestamp
- Bucketing age into categories (0–5, 5–10, 10–15_)
- PricexQuantity= Total_Revenue

FEATURE EXAMPLES - PER TOPIC

FINANCE

- Is_Rounded_Transaction (rounded figures might indicate fraud)
- Percent_Limit_Usage (Current Debt/Current Limit)

HEALTH DATA

- Weight Change (Current weight – previous weight)
- BMI

SUSTAINABILITY

- Weekday vs Weekend energy usage
- Energy_per_square_foot

PRACTICAL PORTION: DATA PREPROCESSING

4. Data Preprocessing & Feature Engineering: Urban Weather Analysis

This session builds directly upon the data initially pulled: the **Location** dataset (containing city coordinates and population) and the **Weather** dataset (containing atmospheric conditions).

Our objective is to transform these raw inputs into a refined, model-ready dataset designed to predict the `feels_like` temperature. Throughout this notebook, we will demonstrate professional data handling techniques including:

- **Relational Merging:** Unifying geographic and atmospheric data points.
- **Leak-Proof Feature Engineering:** Creating a "Wind-Moisture Interaction" index while strictly avoiding data leakage from our target variable.
- **Categorical Encoding:** Utilizing One-Hot Encoding to transform qualitative weather conditions into machine-readable formats.
- **Feature Scaling:** Applying Standardization (Z-score scaling) to the population data to handle variance and the Urban Heat Island effect.

[Link to the Google Colab - Data Preprocessing]

REMINDER

READING WEEK

PERFECT TIME TO CATCH UP ON WORK AND EXPERIMENT

DEADLINE: SESSION 4 (24 FEBRUARY)

- Clear Problem Statement
- Data Chosen and Cleaned
- Ideas for final product in mind

NEXT SESSION

Topic: Machine Learning

Presenter: Rachel + Liam

What We'll Cover

Simple ML Models: Linear Regression, Decision Trees

Advanced: Time-Series, Deep Learning, Ensembles

What to Prepare

💻 Bring your laptop

📌 Reminder of upcoming deliverables

Goal of Session 4

Understand Predictive Modelling and Machine Learning Basics

HAVE A GREAT READING WEEK