BUILDING A DATA SCIENCE PROJECT

UCL DATA SCIENCE SOCIETY
PROJECT LABS 2026

SESSION 1 | 27 Jan 2026

PROJECT PROPOSAL

BE READY TO PITCH YOUR IDEA TO YOUR MENTOR DURING SESSION 2

DEADLINE: SESSION 2 (2 FEBRUARY)

- Formed Groups
- Project Idea
- Problem Statement

* Future changes are expected and acceptable

WHAT IS A DATA SCIENCE PROJECT

- Using Statistics, Programming, Mathematics, and Machine Learning knowledge
- Converting raw data into meaningful insights
- Able to communicate your findings to solve real-world challenges and support decision makings

PROJECT PIPELINE

Step 1:
Problem Definition
(5%)

Step 2:
Collecting data
(10%)

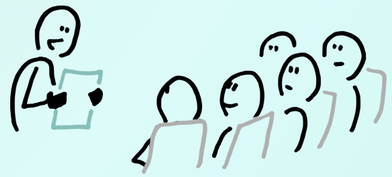
Step 3:
Data Preprocessing
(15%)

Step 4:
EDA
(20%)

Step 5:
Predictive Modelling
(20%)

Step 6:
Visualization &
Interpretation
(20%)

Step 7:
Presentation
(10%)



STEP 1A: PROBLEM DEFINITION – IDEATION

Purpose

Identify **meaningful, feasible,** and **impactful** ideas worth investigating

Key Activities

- Explore domain context (financial, healthcare, climate)
- Review reports, prior analyses, or discuss with peers
- Identify gaps, inefficiencies, or unanswered questions
- Access feasibility (data availability)

Outcome

High-value ideas that are addressable with data

STEP 1B: PROBLEM DEFINITION – PROBLEM FORMULATION

Purpose

Translate an idea into a **clear, actionable** data science question

Key Activities

- Define the analytical objective (prediction, classification, inference)
- Specify inputs (features) and outputs (target)
- Choose evaluation metrics

Outcome

A precise modelling problem ready for data preparation

STEP 1C: PROBLEM DEFINITION – HYPOTHESIS GENERATION

Purpose

Formulate assumptions that **guide feature selection, EDA, and modelling**

Key Activities

- Define expected relationships between features and target
- State null and alternative hypotheses (where applicable)
- Align hypotheses with validation and testing strategies

Outcome

Testable hypotheses that direct EDA and model design

STEP 2: DATA SOURCING & COLLECTION

Data Sourcing

- Identifying relevant data source (database, API, files, and etc)
- Determining whether it is accessible and feasible
- Evaluate data relevance and limitations

Data Collection

- Extract data from identified sources
- Ensure appropriate permissions and ethics
- Store data in a structured and reproducible format

STEP 3: DATA PREPROCESSING

Purpose

Tranforming **raw, real-world** data into a **clean, structured,** and **reliable** dataset

Importance

- Real-world data is **messy** and **inconsistent**
- Models require **structured, valid,** and **numerical** inputs

Outcome

A model-ready dataset with **improved quality and reliability**

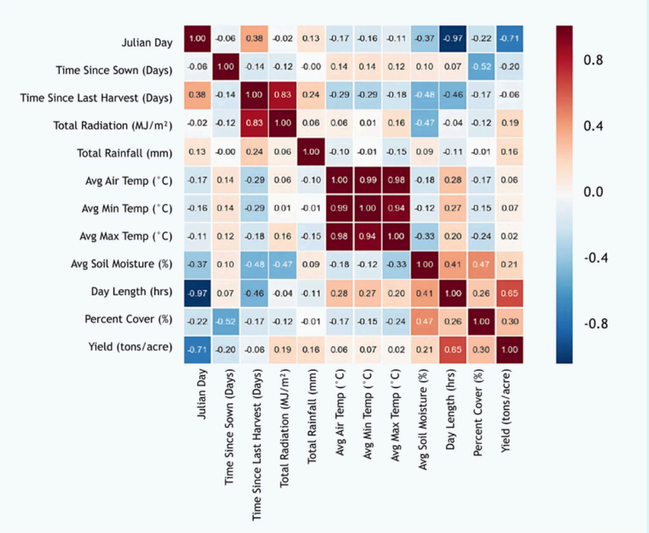
STEP 4: EXPLORATORY DATA ANALYSIS (EDA)

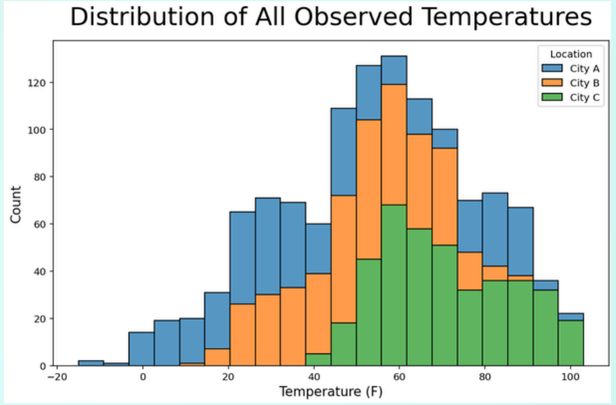
Purpose

- Systematically exploring a dataset to **understand its structure, patterns, and relationships** before modelling
  - Explore **distributions, trends,** and **relationships**
  - Use summary **statistics** and **visualisation**
  - Identify **anomalies, patterns,** and **assumptions**

Outcome

- Data-driven understanding that informs modelling choices





STEP 5: PREDICTIVE MODELLING

Purpose

Build models that **learn patterns** from data to **make predictions or estimates**

Key Activities

- Select appropriate algorithms (e.g. regression, tree-based, ML models)
- Train models using cleaned and engineered features
- Perform validation and hyperparameter tuning
- Evaluate performance using defined metrics

Outcome

A validated model with **quantified performance** and **limitations**

STEP 6: VISUALISATION & INTEPRETATION

Purpose

Turn **EDA findings** and **model results** into **understandable and actionable insights**

Key Activities

- Visualise data patterns discovered during EDA
- Visualise model predictions, errors, and performance
- Interpret feature importance and variable relationships
- Compare findings against expectations or baselines

Outcome

Clear insights explaining **what the data shows, what the model does, and why**

STEP 7: PRESENTATION

Purpose

Communicate findings effectively to technical and non-technical audiences

Key Activities

- Summarise problem, data, approach, and results
- Present insights using clear visuals and narratives
- Highlight assumptions, limitations, and future improvements
- Provide actionable recommendations or conclusions

Outcome

A **clear, credible,** and **decision-ready** project delivery

PROJECT PLANNING & PROGRESS TRACKING

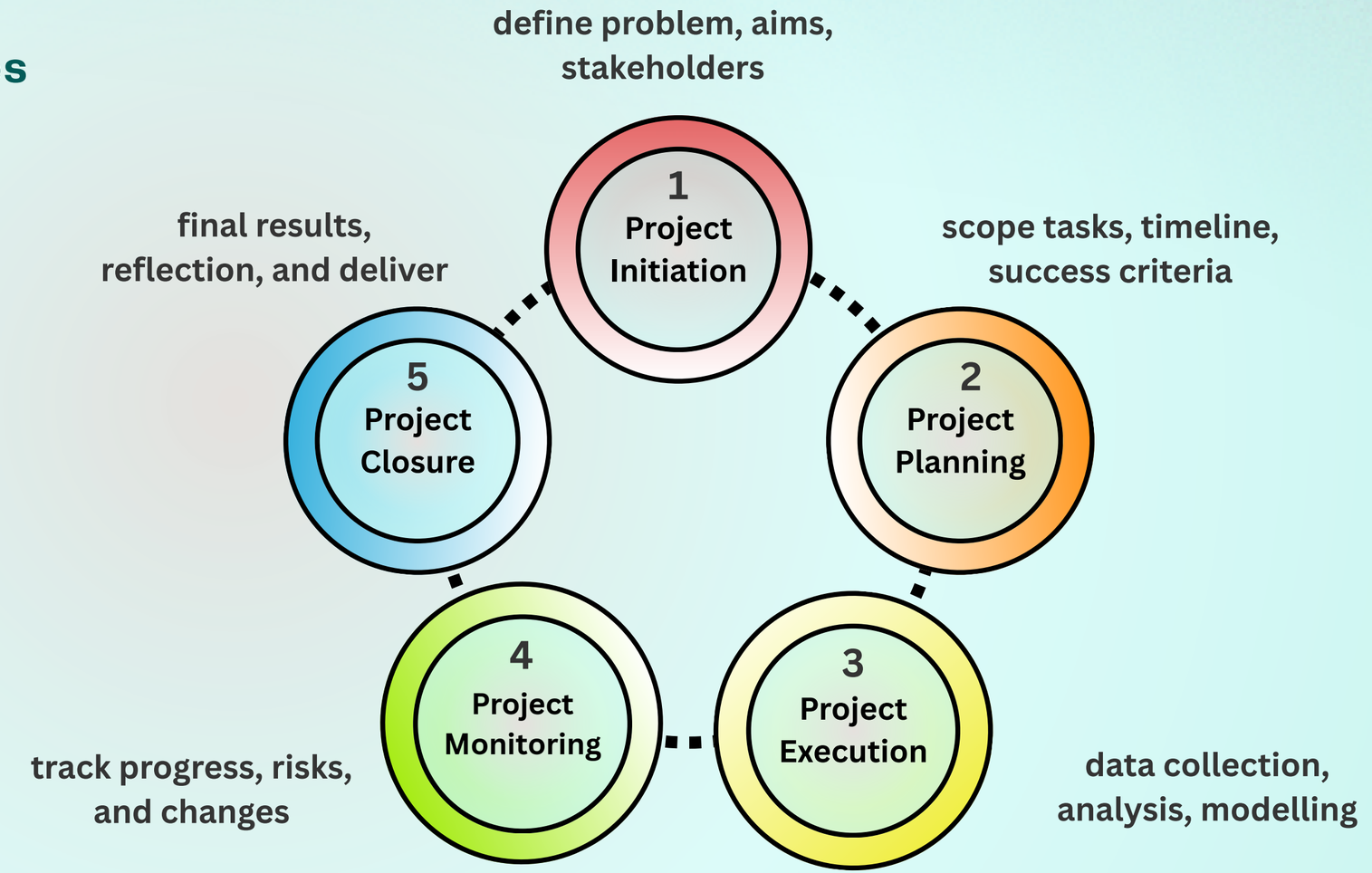
Provide a structured plan that **guides the project** and **supports weekly progress monitoring**

Tools

- Asana
- Trello
- Notion
- **Waiting for your own discovery!!**

Outcome

- A living project plan that evolves while maintaining direction



TOPIC 1: FINANCE

Potential Projects:

- Predicting next-day Amazon stock returns using historical price, volume, and technical indicators
- Detecting fraudulent credit card transactions on anonymized Visa transaction data
- Compare salary growth rates across major industry sectors over time

Data Quality

Transparency & Interpretability

Translation of findings into profit

TOPIC 2 : HEALTHCARE

Potential Projects:

- Risk factor analysis
  - Identify factors associated with a health outcome
- Public health trends analysis
  - Understand population-level patterns and changes over time
- Patient outcome prediction
  - Predict future outcomes using patient-level data

Risk Factor Analysis

- Are patients from lower socioeconomic groups at higher risk of hospital readmission?

Public Health Trends Analysis

- How have hospital admission rates changed over time?

Patient Outcome Prediction

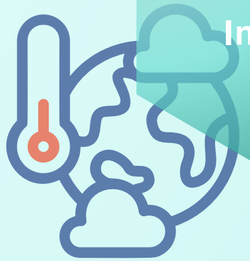
- Can we predict frequent emergency department use?

TOPIC 3 : SUSTAINABILITY & CLIMATE

Multi-Modal Data Sourcing

Interpretability

Measuring Impact



Project Ideas:

- Track and forecast energy consumption used by building when they are supposed to be empty!
- Mapping Air Quality in different parts of London

NEXT SESSION

Topic: Data Sourcing & Preliminary EDA

Presenter: Rachel

What We'll Cover

Data Sourcing

Preliminary EDA

Data visualisations

What to Prepare

💻 Bring your laptop

📌 Reminder of upcoming deliverables

👥 Find and confirm your project group

Goal of Session 2

Leave with a candidate dataset and a first understanding of its structure