DATA SOURCING & PRELIMINARY EDA

UCL DATA SCIENCE SOCIETY
PROJECT LABS 2026

SESSION 2 | 3 Feb 2026

INDICATORS OF A GOOD PROJECT

KEY CHARACTERISTICS

PROBLEM CLARITY

- Is the problem clearly defined?
- Are the data and models clearly aligned to address the problem?

INTERPRETABILITY

- Can the results be clearly explained and translated into insights?

OVERALL COHERENCE

- Does the project flow logically?
- Is the project clear and understandable to others?

PROJECT PIPELINE

Where are we at

Step 1:
Problem Definition
(5%)

Step 2:
Collecting data
(10%)

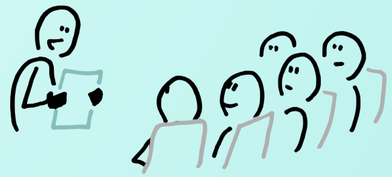
Step 3:
Data Preprocessing
(15%)

Step 4:
EDA
(20%)

Step 5:
Predictive Modelling
(20%)

Step 6:
Visualization &
Interpretation
(20%)

Step 7:
Presentation
(10%)



DATA SOURCING

Purpose

Identify data that is **relevant**, **accessible**, and **appropriate** for the project question.

Key Activities

- Searching for public datasets
- Reviewing the dataset description and variable definitions
- Understanding variables, time range, and granularity

Outcome

A dataset (or set of datasets) that is appropriate for the chosen problem.

DATA SOURCING CONSIDERATIONS

Reliability

- Is the data publicly available and usable?
- Is it trustworthy?

Suitability

- Does the data match the problem statement?
- Does it contain the variables needed to answer the problem?

Feasibility

- Is the dataset size and complexity manageable within the timeline?

STATIC DATASETS

Definition

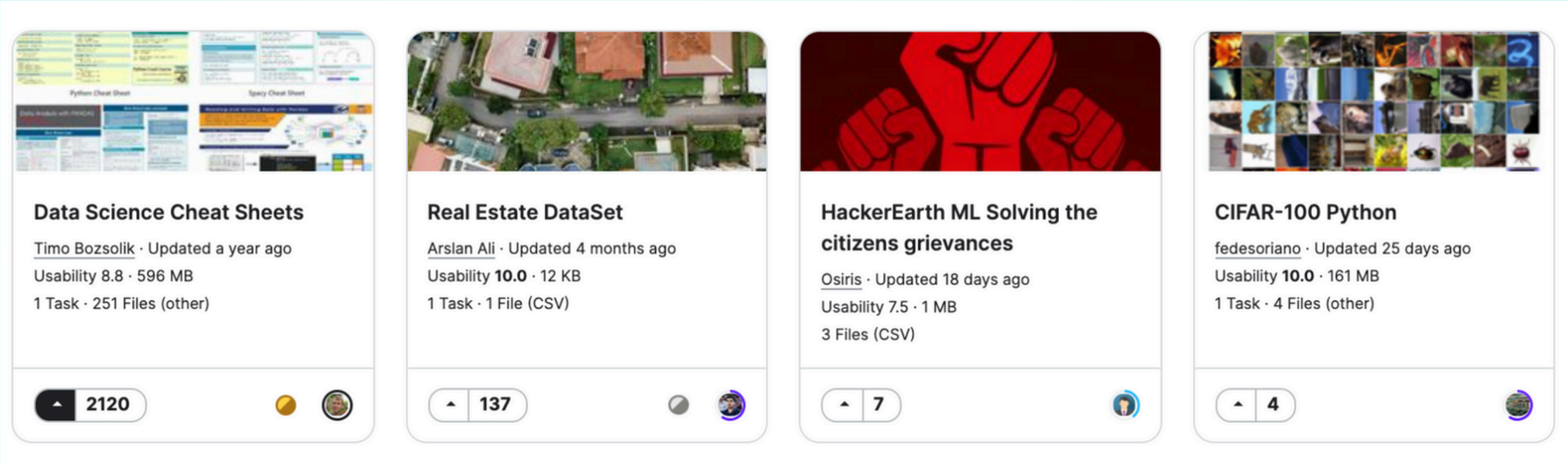
Downloadable datasets (e.g. CSV, Excel) that represent a fixed snapshot of data

Benefits

- Beginner-friendly and easier to work with
- More predictable and reproducible
- Allow you to focus on problem formulation, preliminary EDA, and modeling

Example:

- **Kaggle Datasets**



API & WEB SCRAPING

PURPOSE

- Access and retrieve data that is not readily available as static datasets by programmatically interacting with APIs or web sources.

KEY ACTIVITIES

- Discover and authenticate with an API
- Programmatically request and extract data
- Parse and clean retrieved data
- Store extracted data in structured formats for downstream analysis

EXAMPLES

- Amazon Machine Learning API
- IBM Watson Discovery API

Note: Code for web scraping will not be covered in Project Labs

GENERAL RECOMMENDATIONS

Our World in Data

Research and data to make progress against the world's largest problems.

kaggle

DATA.GOV

UCI Machine Learning Repository

Discover datasets around the world!

ics.uci.edu

THE WORLD BANK

IBRD • IDA | WORLD BANK GROUP

https://www.data.gov.uk

RECOMMENDATIONS - FINANCE

https://data.nasdaq.com/institutional-investors

https://data.ecb.europa.eu

Financial News Sentiment

FRED® ECONOMIC DATA | ST. LOUIS FED

S&P 500 stock data

bitcoin

https://www.alphavantage.co/documentation/

https://pypi.org/project/yfinance/

https://docs.coingecko.com/?utm_source=chatgpt.com

https://datahelpdesk.worldbank.org/knowledgebase/articles/889392-about-the-indicators-api-documentation?utm_source=chatgpt.com

API SERVICES

RECOMMENDATIONS - HEALTHCARE

World Health Organization

uk biobank

HealthData.gov
the home of HHS Open Data

CDC
U.S. CENTERS FOR DISEASE CONTROL AND PREVENTION

OpenNEURO

RECOMMENDATIONS - SUSTAINABILITY

CLIMATE TRACE

NASA EARTHDATA
OPEN ACCESS FOR OPEN SCIENCE

Our World in Data
Research and data to make progress against the world's largest problems.

kaggle

THE WORLD BANK
IBRD • IDA | WORLD BANK GROUP

NOAA

NESO
National Energy System Operator

https://openweathermap.org/api

https://environment.data.gov.uk/apiportal

API SERVICES

Climate Data Store

Google colab

Link to Github

A cloud-based notebook that lets you write and run Python code in your browser.

Basic workflow:

- Open a notebook → write code in cells → run cells → view outputs.
- Mix code with text (Markdown) to explain your thinking.

Benefits:

- No Installation required
- Free GPU Access
- Seamless collaboration

Note: All Project Labs code will be provided in Google Colab notebooks. Please ensure you have access.

PRACTICAL PORTION: CODE FOR DATA SOURCING

1. Data Sourcing & Integration

In this section, we demonstrate how a Data Scientist pulls data from two distinct sources:

1. **Local Data Sourcing:** Internal "proprietary" data (e.g., City Population).
2. **Live API:** Real-time data enrichment (Live Weather).

1.1 Local Data Sourcing (Automated)

Instead of manual uploads, we pull our "internal" population data directly from our GitHub repository. This ensures all members are working with the same version of the data.

- **Primary Data Source:** The underlying dataset is the **World Cities Database**, originally sourced from Kaggle.
- **Storage Method:** Due to the large file size (3.1+ million rows), the CSV is hosted using **Git LFS (Large File Storage)**. This allows us to bypass GitHub's standard file limits and stream the data directly into our environment via Raw GitHub URLs.
- **Automation:** By using `pd.read_csv()` on the hosted URL, we eliminate the need for users to download local copies or manage large files manually.

Column	Description
city	Standardized city name (lowercased for merging)
pop	Total population count
lat / lng	Geographic coordinates (Latitude and Longitude)

Link to the Google Colab – Data Sourcing & Intergation

Google colab

SUMMARY STATISTICS

Purpose

- Describe and summarise the **main features** of a dataset.
- Provide insight into the dataset's **distribution** and **variability**.

Importance

- Helps understand how the data is distributed.
- Shows how spread out or variable the data is.
- Identifies unusual or extreme values (outliers).

--- Numerical Statistics (Weather) ---

	temp	feels_like	humidity	pressure	wind
count	492.000000	492.000000	492.000000	492.000000	492.000000
mean	13.214370	12.063618	66.819106	1017.063008	3.398191
std	13.896821	15.910123	21.708535	9.007922	2.187702
min	-23.930000	-30.460000	6.000000	994.000000	0.000000
25%	3.507500	0.137500	53.750000	1012.000000	1.897500
50%	15.070000	14.130000	70.000000	1016.000000	3.015000
75%	25.302500	25.742500	83.000000	1021.000000	4.640000
max	38.650000	39.950000	100.000000	1042.000000	12.520000

PRELIMINARY EDA

Purpose

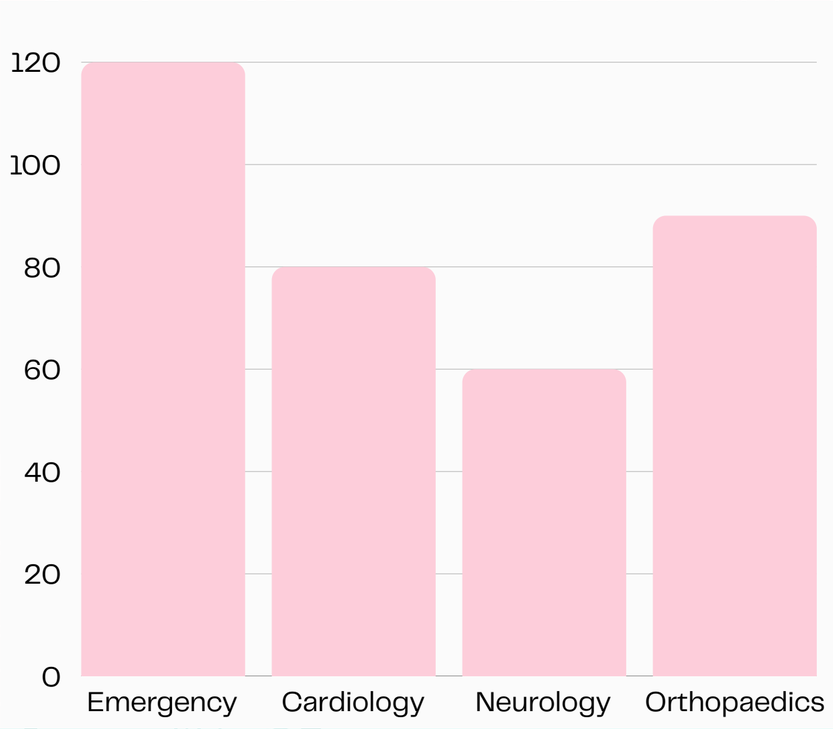
- Initial understanding of the dataset's **structure, quality** and **key characteristics**.
- Detect anomalies and uncover potential patterns or relationships.

Key Activities

- Examine dataset dimensions, data types, and overall structure.
- Generate simple visualisations to explore distributions and relationships.
- Identify outliers, missing values and other data quality red flags.

SIMPLE DATA VISUALIZATIONS

Translate data into visuals to help identify trends, patterns, relationships, and potential outliers.



BAR CHARTS

- Compare values across categories

LINE CHARTS

- Track trends or changes over time

CORRELATION HEATMAP

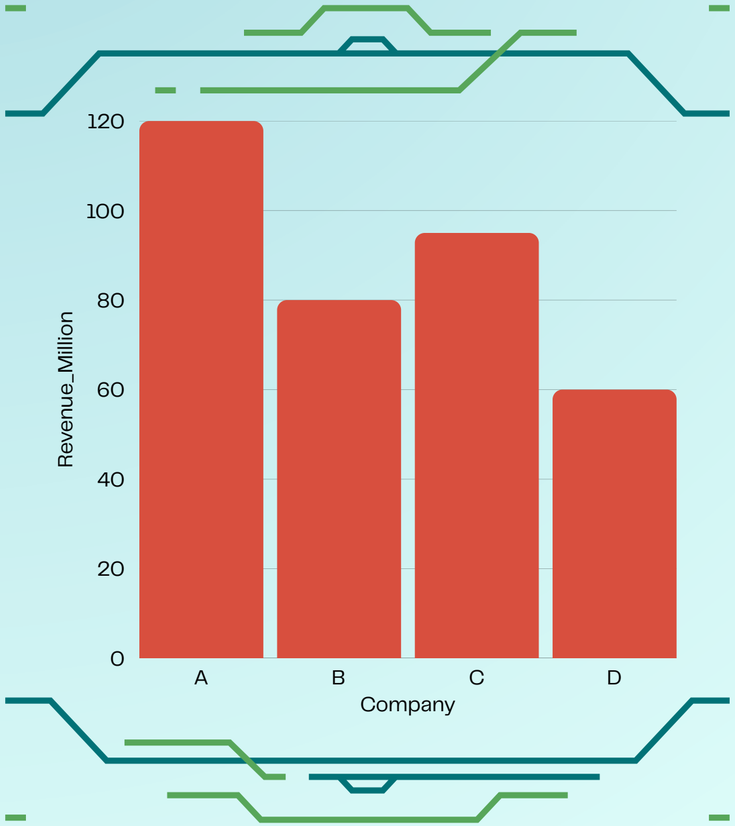
- Explore relationships between multiple numerical variables

SCATTER PLOT

- Examine relationship between variables

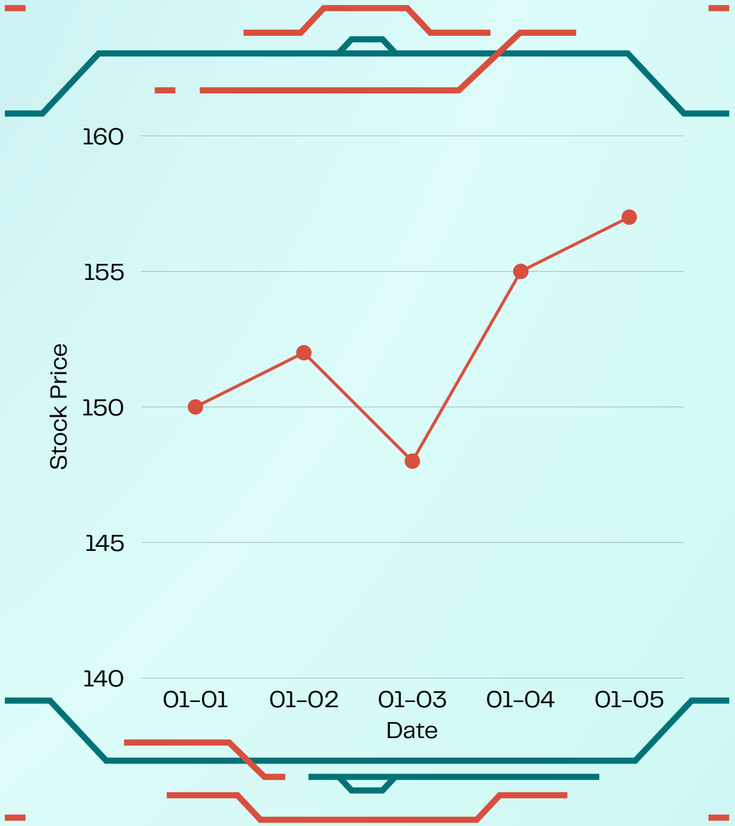
TYPICAL VISUALIZATIONS - FINANCE

Bar Chart



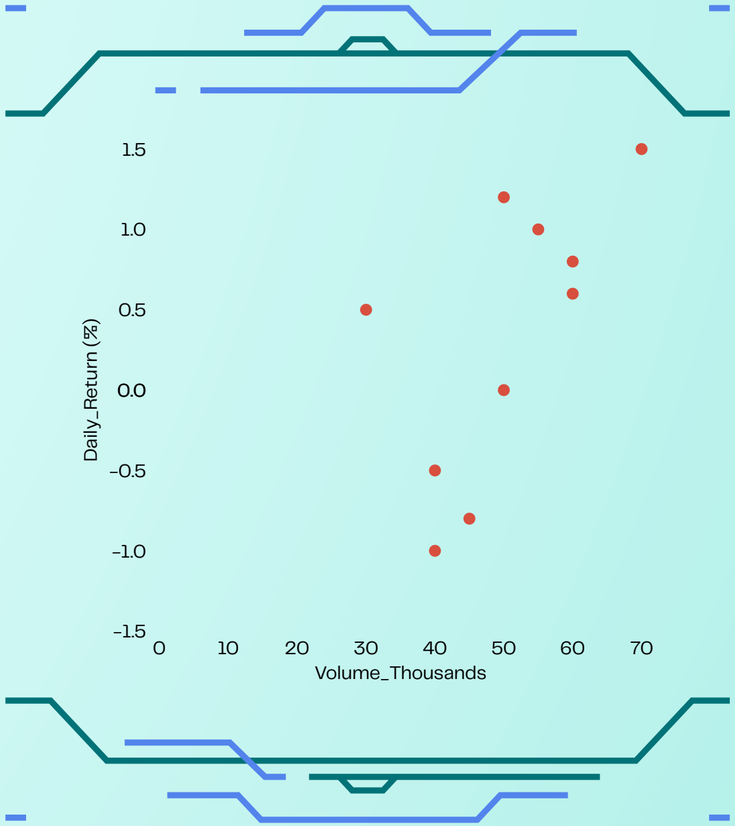
Compare categories, e.g. revenue by company

Line Charts



Tracks trends over time, e.g. stock prices

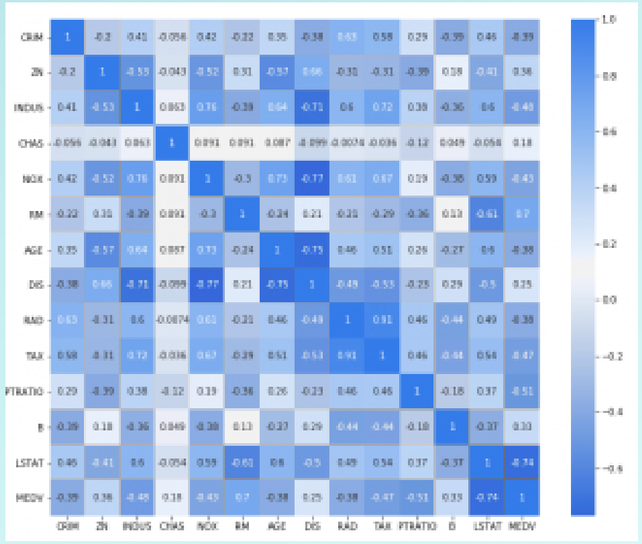
Scatter Plot



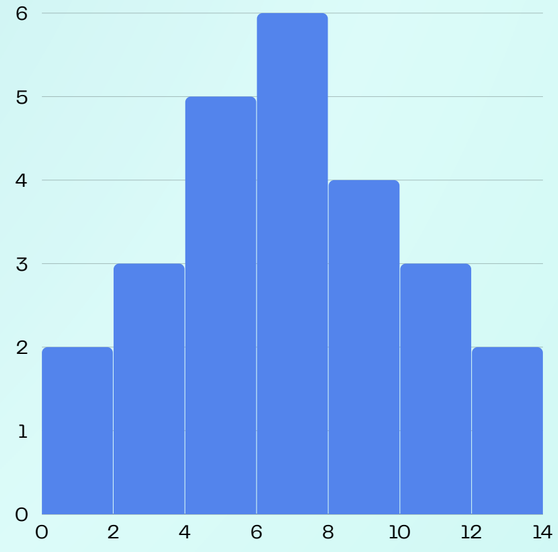
Examine relationship between 2 numeric variables

TYPICAL VISUALIZATIONS - HEALTHCARE

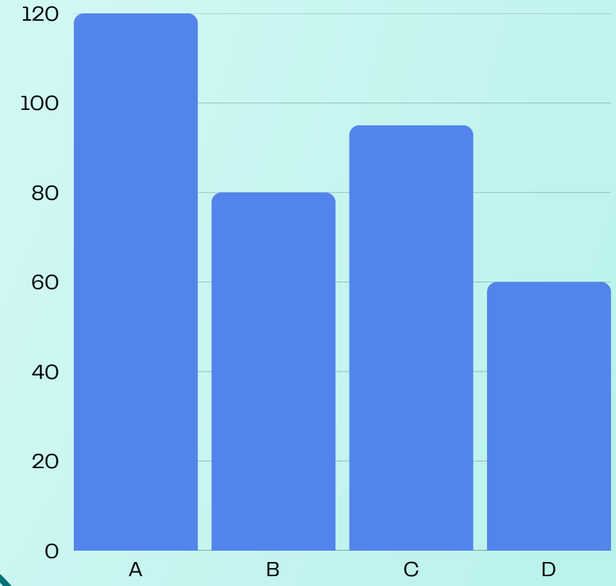
Correlation Heatmap



Distributions – Histograms/Boxplots

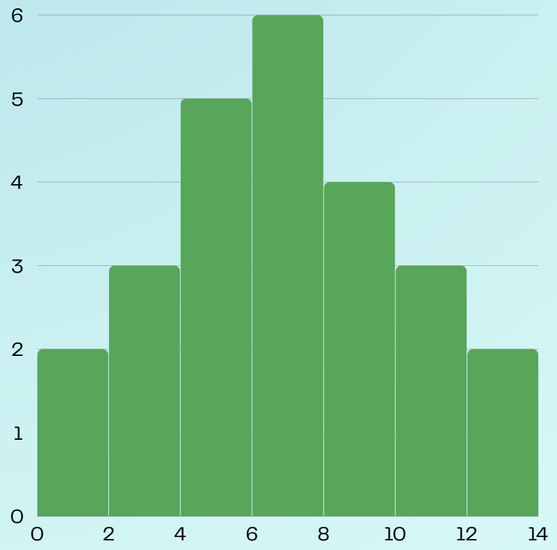


Missingness Visualisation

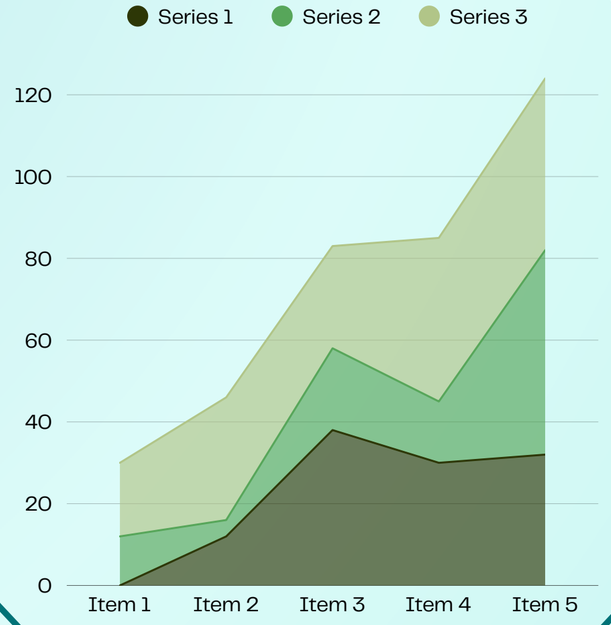


TYPICAL VISUALIZATIONS - SUSTAINABILITY

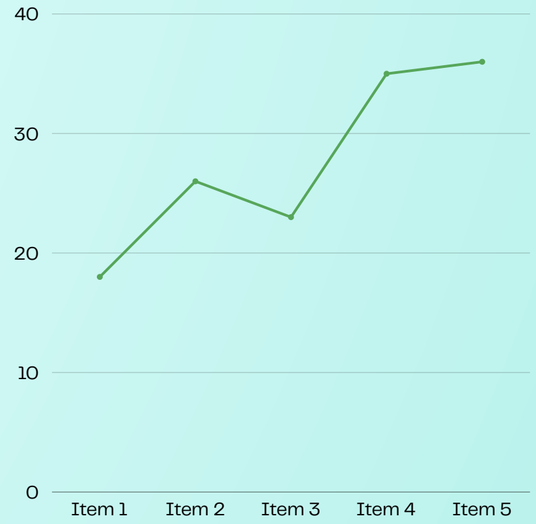
Histograms



Stacked Area Charts



Line Graphs



‘BIG PICTURE’ ANALYSIS (PAIRPLOTS)

Purpose

Provide an overview of how multiple variables relate to one another.

Key Aspects

Matrix of plots

- Display pairwise relationships

Diagonals

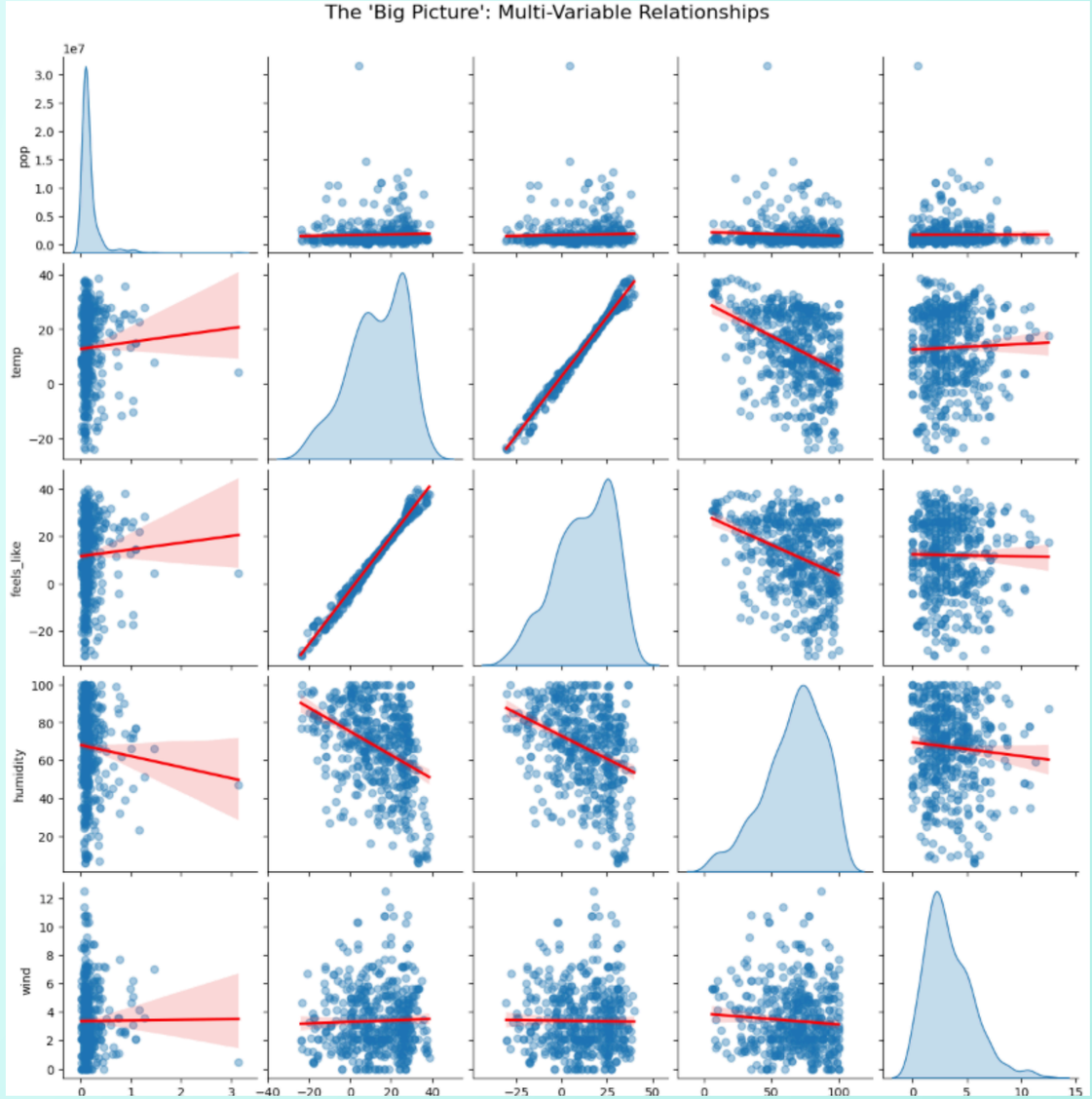
- Distribution of each variable

Scatter Plots

- Relationship between pairs of variables

Regression Line(s)

- Highlight potential linear trends



PRACTICAL PORTION: CODE FOR PRELIMINARY EDA

2. Preliminary EDA & Visualization

Goal: Understand the "shape" and "health" of the data.

```
# --- LFS SETUP & REPO CLONING ---
import os
import pandas as pd

# 1. Install Git LFS in the Colab environment
!git lfs install

# 2. Clone the repository (This pulls the LFS pointers)
REPO_NAME = "DS-Society-Project"
REPO_URL = f"https://github.com/LiamDuero03/{REPO_NAME}.git"

if not os.path.exists(REPO_NAME):
    !git clone {REPO_URL}
else:
    print("✅ Repository already cloned.")

# 3. Explicitly pull the LFS data (This replaces pointers with actual CSV data)
%cd {REPO_NAME}
!git lfs pull
%cd ..
```

Link to the Google colab – Preliminary EDA

Google colab

REMINDER

TECHNICAL PROPOSAL

BE READY TO SHARE YOUR IDEA DURING SESSION 3

DEADLINE: SESSION 3 (10 FEBRUARY)

- Outline Document
  - Chosen dataset(s)
  - Clear problem statement
  - Summary statistics

NEXT SESSION

Topic: Project Tools & Data Wrangling

Presenter: Liam

What We'll Cover

Github/Git and Project Management Tools

Data Wrangling

Preparing Data for Modelling

What to Prepare

💻 Bring your laptop

📌 Reminder of upcoming deliverables

Goal of Session 3

Leave with a cleaned data base and a shared github repository