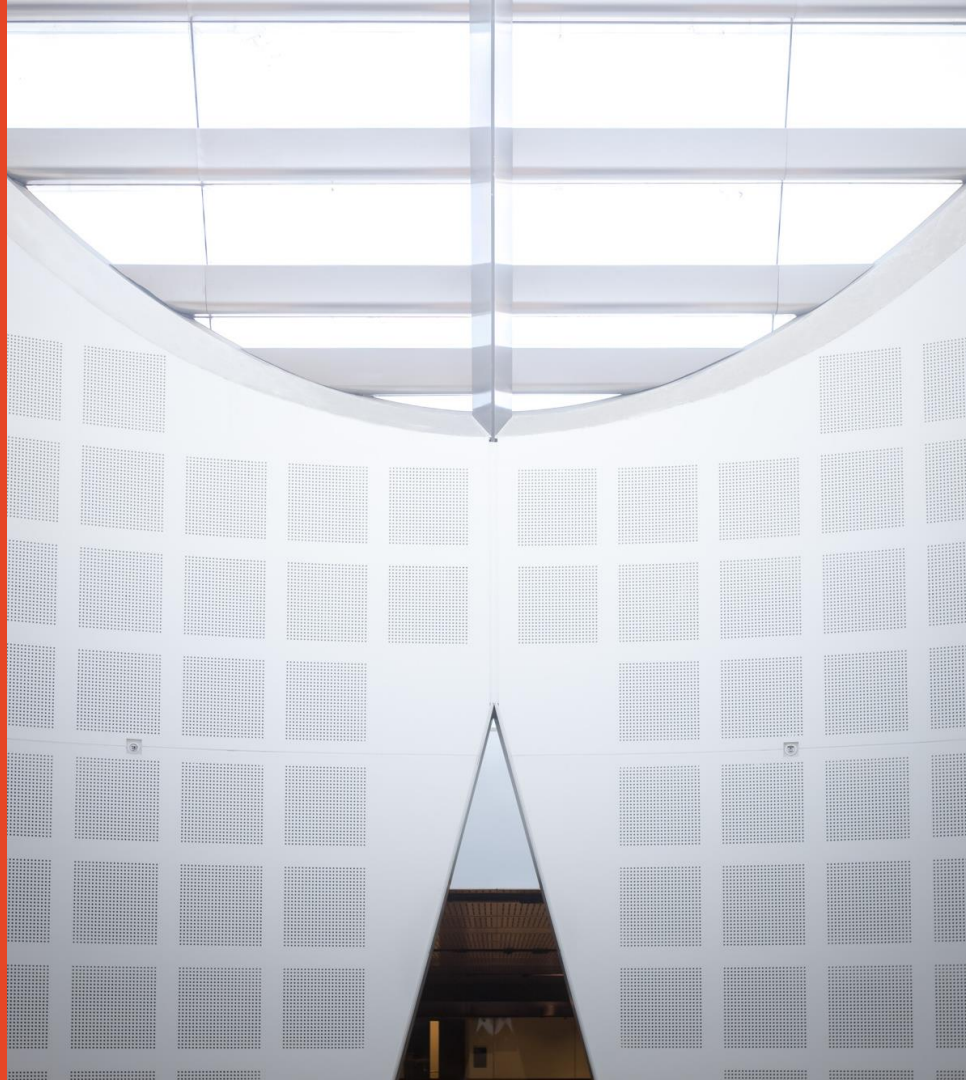
OCMP5310: Principles of Data Science

Week 2 Live Session

Presented by
Daniela Rivas

THE UNIVERSITY OF SYDNEY



LIVE SESSIONS

THE UNIVERSITY OF SYDNEY

Live Sessions Details

- 6 Weeks.
- Every **Monday** at **6:30 PM**.
- First Live Session: Monday 17th of April.
- Last Live Session: Monday 22nd of May.
- 90 minutes.

Live Sessions Expectations

Before the session:

- Watch the lecture videos.
- Work on the practical exercises.

During the session:

- Reflect on the main topics of the week.
- Run some additional group exercises.
- Give feedback on the exercises and the assessment tasks.

REVIEW OF LAST WEEK

Week 1

- Installing Python and PostgreSQL.
- Exploratory Data Analysis with Spreadsheets/Python (Calculating Descriptive Statistics, plotting).
- Descriptive Statistics and Data visualization with Python.

QUESTIONS?

WEEK 2: SESSION ACTIVITY

THE UNIVERSITY OF SYDNEY

Week 2

- Creating/querying databases with SQL.
- Analysing/summarising data from multiple Tables (joining) with SQL.

Creating/querying databases with SQL

Activity

- In Canvas, go to:
  - Exercise: Data storage with Python.
- Download Jupyter Notebook:
  - ETL_with_python_and_postgres.ipynb
- Download files:
  - Measurements.csv
  - Organisations.csv
  - Sensors.csv
  - Stations.csv
  - Credentials.json

Analysing/summarising data with SQL

Activity

- In Canvas, go to:
  - Exercise: Summarising data with SQL.
- Download Jupyter Notebook:
  - Summarising_Data_with_SQL.ipynb

Project Stage 1

Project Stage 1 (10% overall mark)

- **Due:** Week 3, 7 May 2023, 23:59, Sydney Time.
- **Tasks:**
  - Selecting a dataset.
  - Defining the problem and project requirements.
  - Acquiring and loading the dataset into either a database or a Jupyter notebook.
  - Data cleaning.
- **Submissions:**
  - 2-page report.
  - Code.
- More details here.

Report

- 2-page report (not counting title page and references or appendix) that describes the problem, proposed approach and dataset, and data cleaning process.
  - **Problem:** Describe the problem from a general perspective, highlighting the business/research need. List the research question(s) you will answer in Stages 2 and 3 of the project.
  - **Approach:** Describe the approach you will take to solving the problem and any requirements. This is your plan for Stage 3.
  - **Data:** Describe the data from a general perspective e.g. source, size, fields of interest. How did you acquire the data? Describe any data preparation steps e.g. transformation, sampling, cleaning.

Code

- Your code used to ingest and clean your dataset.
- Format: Jupyter Notebook (.ipynb), python script (.py) or similar.

QUESTIONS?