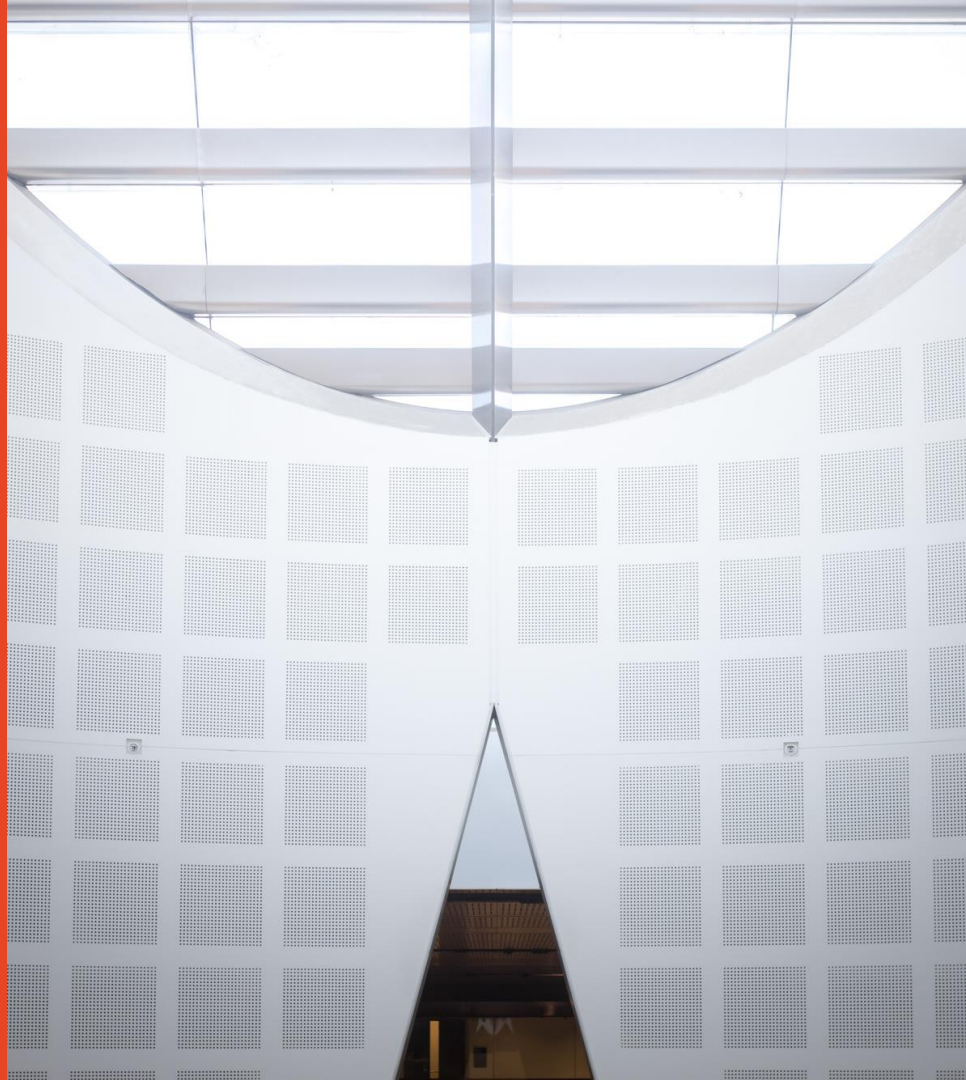
OCMP5310: Principles of Data Science

Week 4 Live Session

Presented by
Daniela Rivas

THE UNIVERSITY OF SYDNEY



REVIEW OF LAST WEEK

Week 3

- Hypothesis Testing.
- Model Evaluation.

QUESTIONS?

WEEK 4: SESSION ACTIVITY

THE UNIVERSITY OF SYDNEY

Week 4

- Unsupervised learning: Clustering.
- Supervised learning: Regression and classification.

Clustering and Dimensionality Reduction

Activity

- In Canvas, go to:
  - Exercise: Clustering Exercise.
- Download Jupyter Notebook:
  - clustering_and_dimensionality_reduction.ipynb

Regression

Activity

- In Canvas, go to:
  - Exercise: Linear Regression.
- Download Jupyter Notebook:
  - regression.ipynb

Decision Trees

Activity

- In Canvas, go to:
  - Exercise: Decision Trees.
- Download Jupyter Notebook:
  - decision_tree.ipynb

Project Stage 2

Project Stage 2 (10% overall mark)

- The purpose of Stage 2 is to **summarise**, and **analyse** the dataset to get a good understanding of your dataset and as a preparation for the final project stage.
- **Due:** Week 5, 16 May 2023, 23:59, Sydney Time.
- **Tasks:**
  - Describe the problem.
  - Overview of data.
  - Exploratory Data Analysis.
- **Submissions:**
  - 2-page report.
  - Code.

Report

- 2-page report (not counting title page and references or appendix) that describes the problem, proposed approach and dataset, and exploratory analysis performed.
  - **Problem:** Describe the problem from a general perspective and list the research question(s) you will answer in Stage 3 of the project.
  - **Data:** Describe the data from a general perspective.
  - **Approach:** Describe any exploratory analysis you have done to refine your understanding of the data and research question. One or two supporting figures/tables would be great.

Code

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

QUESTIONS?