COMP2200/COMP6200 Lecture 1d – Orange Clustering

Greg Baker

23rd February 2026

What is Orange?

- Visual data mining suite built on top of Python
- Drag-and-drop widgets let you build workflows quickly
- Great for exploring data without writing code

https://orangedatamining.com/download/

Downloading Orange

- Visit https://orangedatamining.com/download/
- Choose the installer for your operating system
- It's around 500MB, so start the download now

https://orangedatamining.com/download/

What is Clustering?

Easy Definition

Clustering is a way of grouping similar things together, without being told what the groups should be.

- Imagine sorting socks from the laundry —— you group them by colour or pattern, even if no one told you what the categories are.
- Practical examples:
  - Grouping customers by shopping behaviour (customer segmentation)
  - Detecting communities in social networks
  - Organising news articles into topics

https://orangedatamining.com/download/

Give me a game-ified learning environment

`http://clustering-visualizer.web.app`

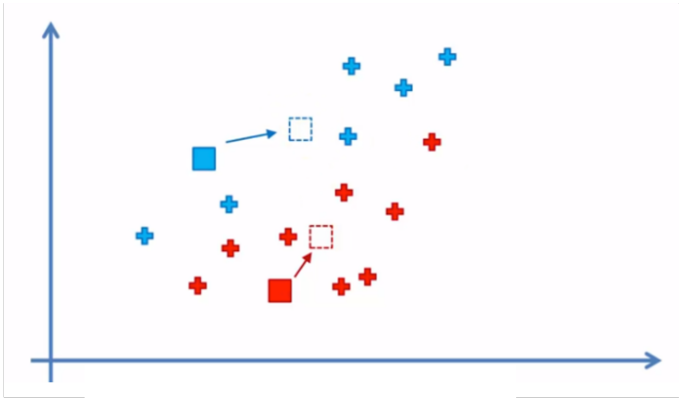
Feel free to ignore the lecture and learn on your own from this one web app.

`https://orangedatamining.com/download/`

K-Means Clustering

How it works

- Choose how many clusters you want (say, $k = 3$).
- Drop k points randomly — these are your cluster centres (called **centroids**).
- Assign each point in the dataset to the closest centroid.
- Move the centroids to the **mean** of the points in their cluster.
- Repeat until nothing changes.



https://orangedatamining.com/download/

Example: Indian Supermarkets

Pick your scenario:

- You have a business supplying something to Indian supermarkets, and you need to think about distribution depots
- You are looking for a place to rent and you want to see what suburbs to consider so that you not too far away from an Indian supermarket?

Q: Why choose Indian supermarkets?

A: the cluster diagram looked nice

https://orangedatamining.com/download/

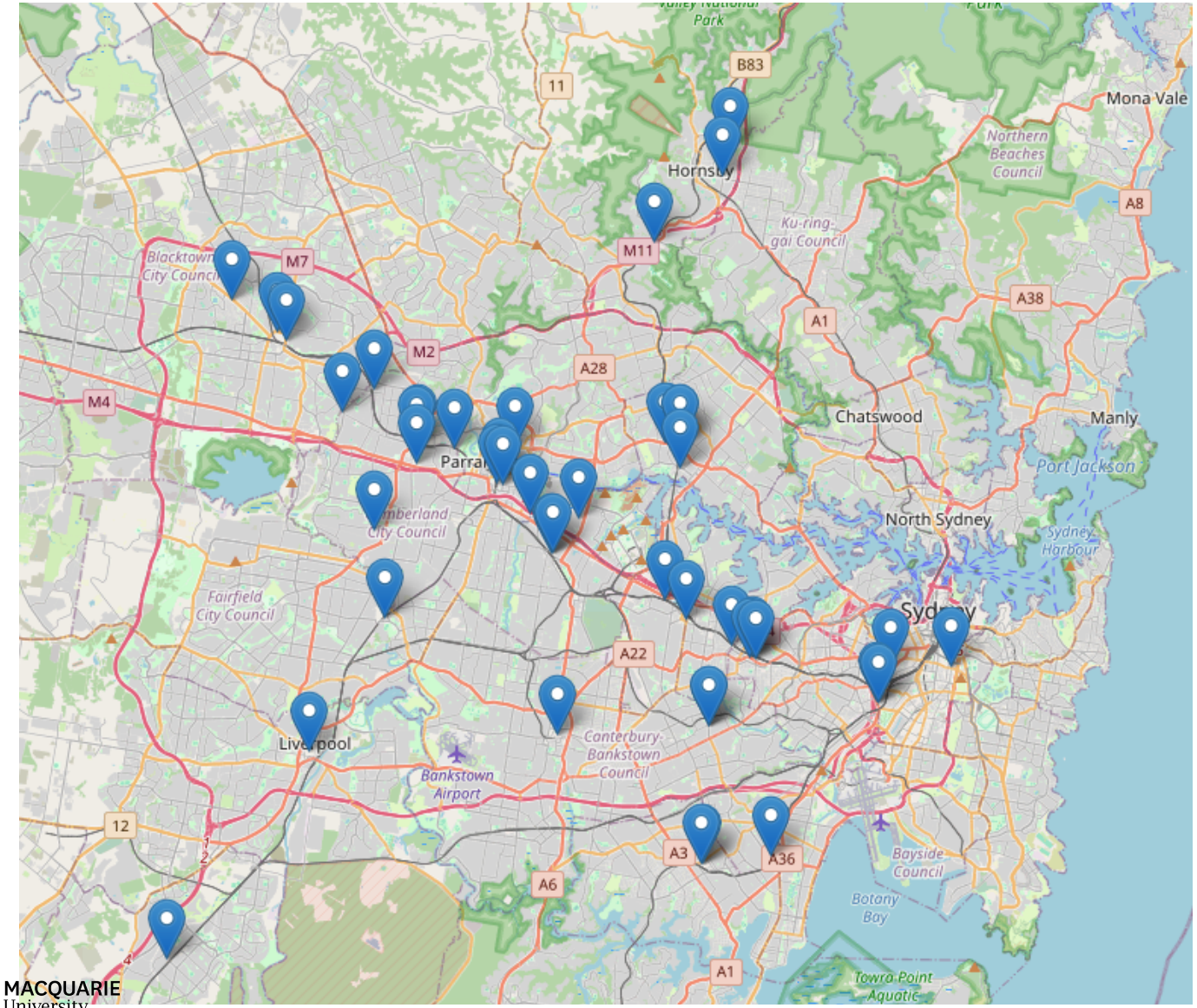
Geocoding

Ways to turn a street address into a (latitude,longitude) pair.

- Search in Google maps, then right click
- Government databases
- Pick a programming API from
  `https://wiki.openstreetmap.org/wiki/Geocoding`
- Google Sheets plug-in

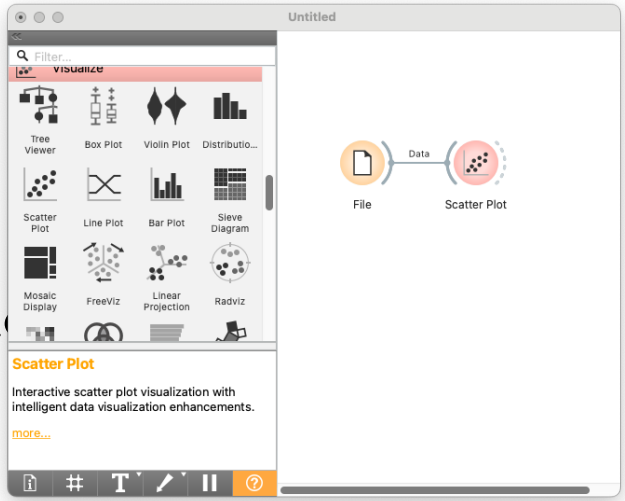
https://orangedatamining.com/download/

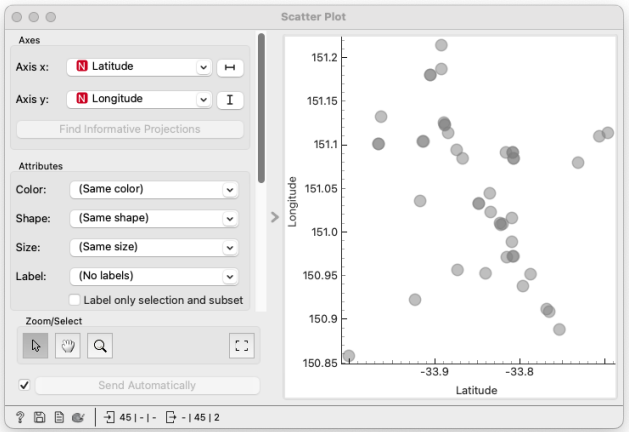
Let's visualise it



Plotting in Orange

1. Start Orange and add a **File** widget.
2. Load `indian-supermarkets-in-sydn`
3. Add a **Scatter Plot** widget and connect it to the File widget.
4. Choose *latitude* for the x-axis and *longitude* for the y-axis.
5. The map of supermarket locations should now appear.





https://orangedatamining.com/download/

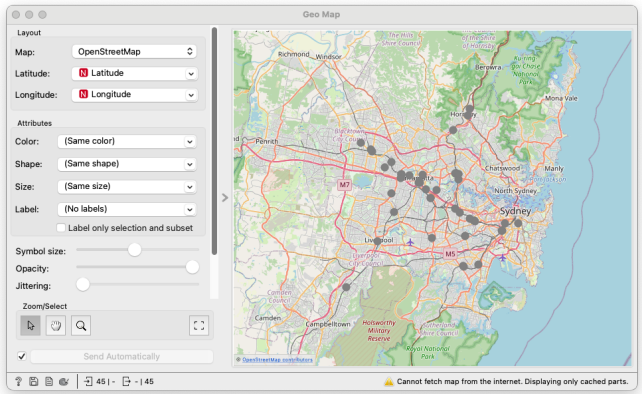
Loading a CSV in Python (pandas.read_csv)

```
import pandas as pd
df =
    pd.read_csv("indian-supermarkets-in-sydney.csv
df.head()
```

- In Google Colab: upload the CSV first (Files pane → Upload).
- A **DataFrame** is a table of rows and columns (think spreadsheet).

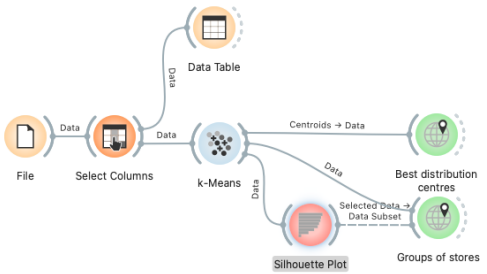
Making it nicer

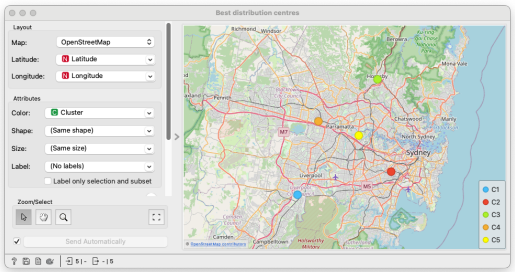
- Select Columns to ignore the null columns, and make the Latitude and Longitude into Features
- Add a Data Table to see the data
- Add the Geo add-in
- Add a Geo Map

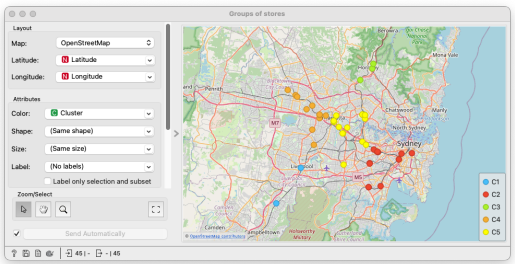


k-Means in Orange

1. Add a **k-Means** widget and connect it to your data. Turn off normalisation
2. Select how many clusters you want to create.
3. Connect the output to a **Scatter Plot** or **Geo Plot** to colour points by cluster.
4. Use the **Silhouette Plot** widget to judge the quality of the clustering.
5. Add another **Scatter Plot** or **Geo Plot** to visualise where the centroids are







What we haven't talked about

- What's that Silhouette Plot doing?
- How do we calculate a Silhouette score?
- Why did we turn off normalisation?

We'll cover these topics in a few weeks' time

Wrapping up

- Clustering helps you find natural groupings in data, even without explicit labels.
- K-Means clustering is simple and powerful: choose k, assign, average, and iterate.
- Orange makes clustering easy with a visual, drag-and-drop approach.
- Geo visualisations enhance insights when clustering geographic data.

Next steps:

- Try clustering your own datasets.
- Explore different values of k.
- Experiment with other widgets in Orange.