CORE Skills

Delivering Data Science
In Resources & Energy

Data Analysis I: Data munging and exploratory data analysis

DAY 4

15-Day Data Science Springboard

Dr Jeremy Mitchell & Dr Ying Yap, Data Mettle

Program partners

_dm

THE UNIVERSITY OF WESTERN AUSTRALIA

Curtin University

A CORE partnership

Program Timeline

DAY 4, 5 & 6: Data Analysis

Leader Engagement	15 Day Professional Program		
2-hour Leading Data Scientists Leader Support	Preparatory		Introduction to Data Projects
2-hours	Day 1	Day 2	Day 3
January 2022	Wed 24 Nov 2021	Thurs 25 Nov 2021	Wed 1 Dec 2021
Enabling your people's data science upskilling & project delivery in 15 day program	Introduction to the Program Tools	Introduction to the Program Tools:	Zero to Data Science in a day

Data Analysis

Day 4	Day 5	Day 6
Wed 8 Dec 2021	Wed 15 Dec 2021	Wed 19 Jan 2022
Getting to know the program tools - data munging and exploratory data analysis	Simple predictions - regression and statistical model building	Multivariate analysis and model building

Data & Communication Sandbox	Data Fusion and Machine Learning		Data Fusion Sandbox	Special Data Types - Time-series & Networks	Special Data Types - Knowledge Discovery from Natural Language Processing & Text Mining	Special Data	Capstone Project Development & Presentation	Capstone Propeller
Day 7	Day 8	Day 9	Day 10	Day 11	Day 12	Day 13	Day 14	Day 15
Wed 2 Feb 2022	Wed 9 Feb 2022	Wed 16 Feb 2022	Wed 23 Feb 2022	Wed 2 Mar 2022	Wed 9 Mar 2022	Wed 16 Mar 2022	Wed 23 Mar 2022	Wed 22 June 2022
Effective data storytelling: Communicating results to non-technical audiences	Pros and cons of commonly used statistical and machine learning techniques I	Pros and cons of commonly used statistical and machine learning techniques II	Consolidate approaches covered and test on datasets	The 4th dimension and predictions	Finding needles in wordstacks	Spatial analytics and predictions	Capstone Project pitches to leadership	Project Review

Q&A, Issues & Announcements

Before we Get Started

- Resources & Tasks in notebooks.
- Where practical we'll be working with *your* data.
- You'll be doing some guided exploration - this is a good time to get in some practise & find some new features of `pandas`, `matplotlib` and `seaborn`.
- Make notes about any ideas, perspectives or issues you encounter throughout the day.
- If you have aspects you'd like to go over throughout the day, feel free to post them to the general channel and we'll try to address the straightforward ones as we break.
- We'll come together to discuss before we close out this afternoon.

Schedule

DAY 4

AWST	AEST	Agenda	Educator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Munging Tabular Data**	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	**Grouping & Reshaping**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	**Explaining Data**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Practice Explaining Your Own Data**	Jeremy
14:45	16:45	**Closeout** – Reflections, Takeaways & Project Selection	Jeremy
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Aims & Learning Outcomes

DAY 4

Aims

- Create and interpret statistics and visualisations after completing appropriate QA/QC.
- Implement exploratory data analysis and apply pattern recognition principles while avoiding pitfalls.

Learning Outcomes

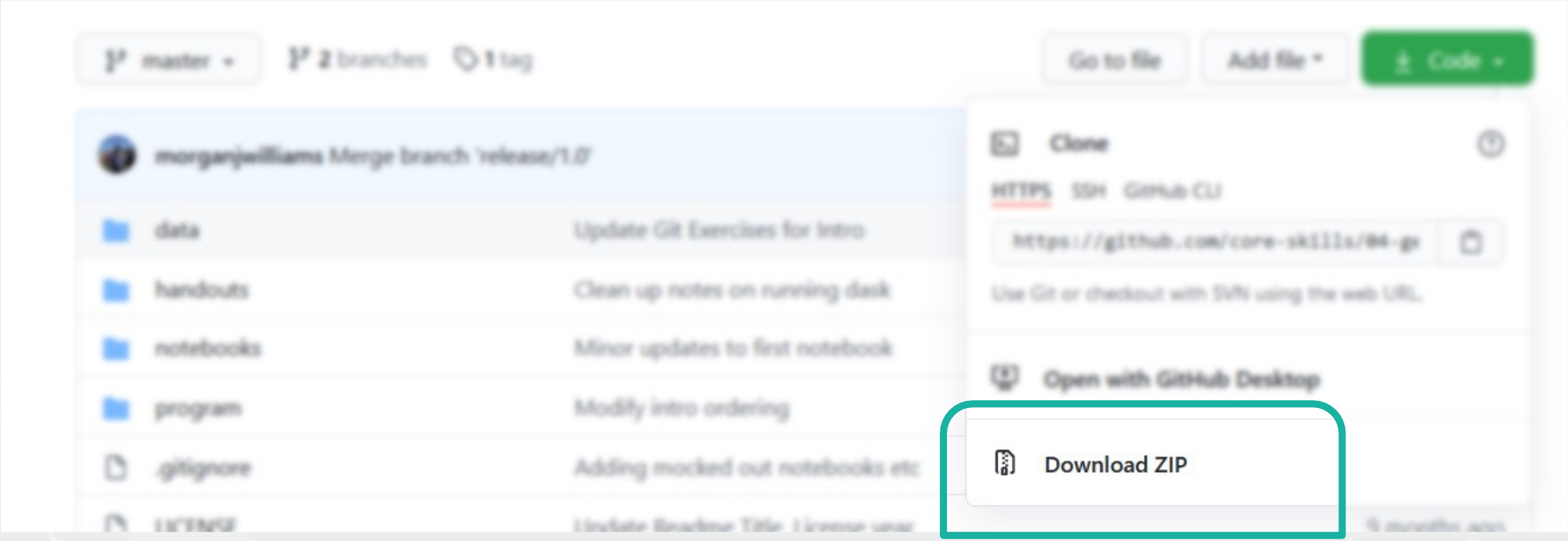
- Understand the pitfalls of different data types.
- Appreciate the importance of choosing i.e. 'clean data' and be aware of some QA/QC approaches for enforcing this.
- Perform basic data visualisations given tabular data.
- Construct reasoning to explain links between data and statistical distributions including pattern recognition.
- Critique basic summary statistics after implementing exploratory data analysis.

Getting to Know the Tools

- Weeks to come will be about building models
- Today is about getting comfortable with the tools we'll continue to use
- The pace is still relatively slow today to get everyone on the same page
- If you finish exercises take the opportunity to dive into the docs and find something new

GitHub Content for Today

[github.com / core-skills / 04-getting-to-know-the-tools]



GitHub – Program Notes

github.com / core-skills / 04-getting-to-know-the-tools / program / 00_overview.md

Overview

Overview | Munging | Grouping & Reshaping | Explaining Data | How Might We... | Closeout

Aim

1. Create and interpret statistics and visualisations after completing appropriate QA/QC.
2. Implement exploratory data analysis and apply pattern recognition principles while avoiding pitfalls.

Learning Outcomes

1. Understand the pitfalls of different data types.
2. Appreciate the importance of choosing i.e. 'clean data' and be aware of some QA/QC approaches for enforcing this.
3. Perform basic data visualisations given tabular data.
4. Construct reasoning to explain links between data and statistical distributions including pattern recognition.
5. Critique basic summary statistics after implementing an EDA.

Schedule

AWST	AEST	Agenda
07:30 - 07:45	09:30 - 09:45	Q&A, Issues & Announcements
07:45 - 09:15	09:45 - 11:15	Munging Tabular Data

Environment

- Open an Anaconda Prompt
- Navigate to where you have the unzipped repository material

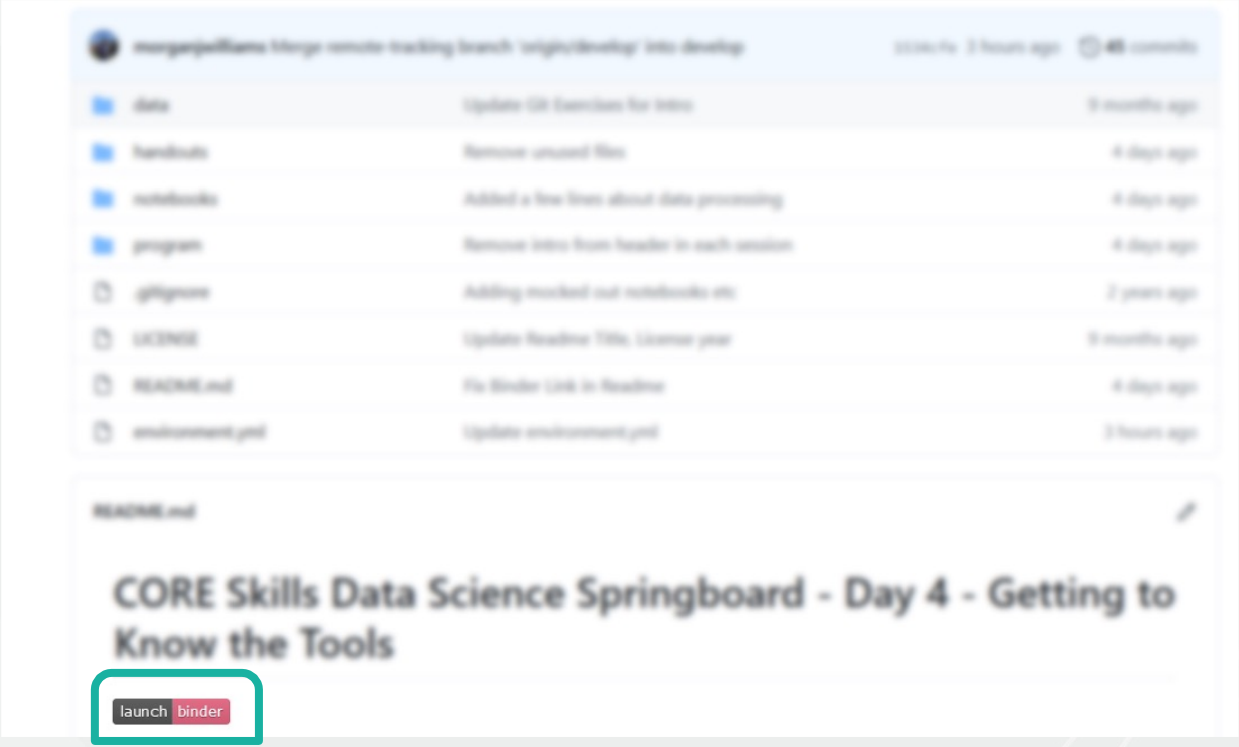
```
conda env create -f environment.yml

conda activate core04

python -m ipykernel install --user --name=core04

jupyter lab
```

Binder Backup



Scientific Python Ecosystem

Package	Description
numpy	Numeric Python. Working with numbers, lists of numbers, linear algebra & more.
matplotlib	"2D plotting library which produces publication quality figures in a variety of formats."
pandas	"A fast, powerful, flexible and easy to use (tabular) data analysis and manipulation tool"
sklearn	"Machine Learning in Python- Simple and effective tools for predictive data analysis"
scipy	Optimisation, interpolation, signal processing & stats.
statsmodels	"for the estimation of statistical models, statistical tests, statistical data exploration"
seaborn	"High-level interface for drawing attractive and informative statistical graphics."

Pair Programming

Suggestion for the Non-Virtual World, or 1-on-1 Video Collaborations:

- Two people work on one computer
- One person writing code, the other reviewing & making suggestions
- This shouldn’t be a quiet exercise – it’s closer to ‘coding out loud’
- A good way to share and consolidate knowledge
- In a development scenario it’s ≈ 15% slower than two people coding, but you end up with significantly better results in terms of code quality

Reading Data (20 mins)

First Exercise – Step 1: Reading Data

We'll run through most of this together.

notebooks/am1-munging-tabular-data.ipynb

Status

Checkpoint

What’s In my Data?

(20 mins)

First Exercise – Step 2: What’s in my Data?

notebooks/am1-munging-tabular-data.ipynb

Status

Checkpoint

Tidy Data (20 mins)

Tidy Data

The Checklist:

- Each variable you measure should be in one column.
- Each different observation of that variable should be in a different row.
- There should be one table for each "kind" of variable.
- If you have multiple tables in a given dataset, they should include a column in the table that allows them to be linked

First Exercise – Step 3

Use the Tidy Data Checklist to Check Your Own Data

notebooks/am1-munging-tabular-data.ipynb

Status

Checkpoint

Optional:
Extract These Steps out to a Separate Function

Schedule

AWST	AEST	Agenda	Facilitator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Munging Tabular Data**	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	**Grouping & Reshaping**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	**Explaining Data**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Practice Explaining Your Own Data**	Jeremy
14:45	16:45	**Closeout** – Reflections, Takeaways & Project Update	Tamryn
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Grouping & Reshaping Data (20 mins)

df.groupby()

Group-by is one way of mapping through a for loop within a DataFrame.

- Split part of **split-apply-combine**
- Especially when you use the aggregation functions
- You can group by multiple columns, but it can get complicated quickly

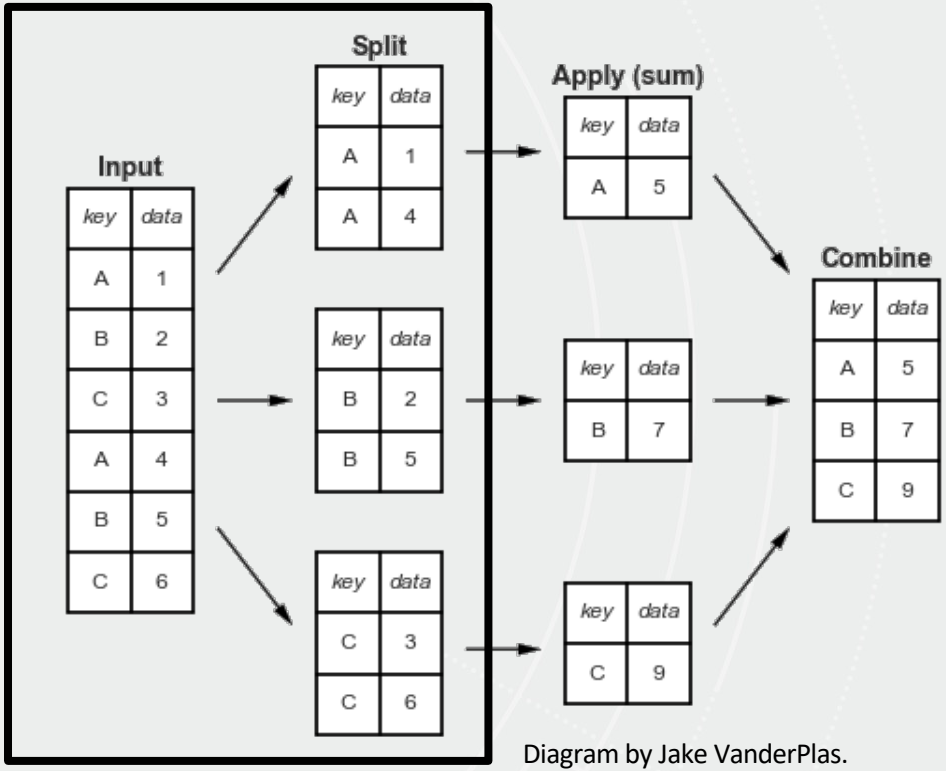


Diagram by Jake VanderPlas.

df.pivot() /pd.pivot_table()

In some cases, pivots can provide similar functionality.

They're useful for **tidying tables** where variables are in rows, but we can also use **aggregation functions** here!

They can be a bit difficult to get your head around to start with.

df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

Pivot

```
df.pivot(index='foo',
         columns='bar',
         values='baz')
```

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

Diagram from Pandas docs.

df.pivot() /pd.pivot_table()

df.pivot() /pd.pivot_table()

When you're using aggregation functions, group-by and pivot can both facilitate the split-apply-combine approach.

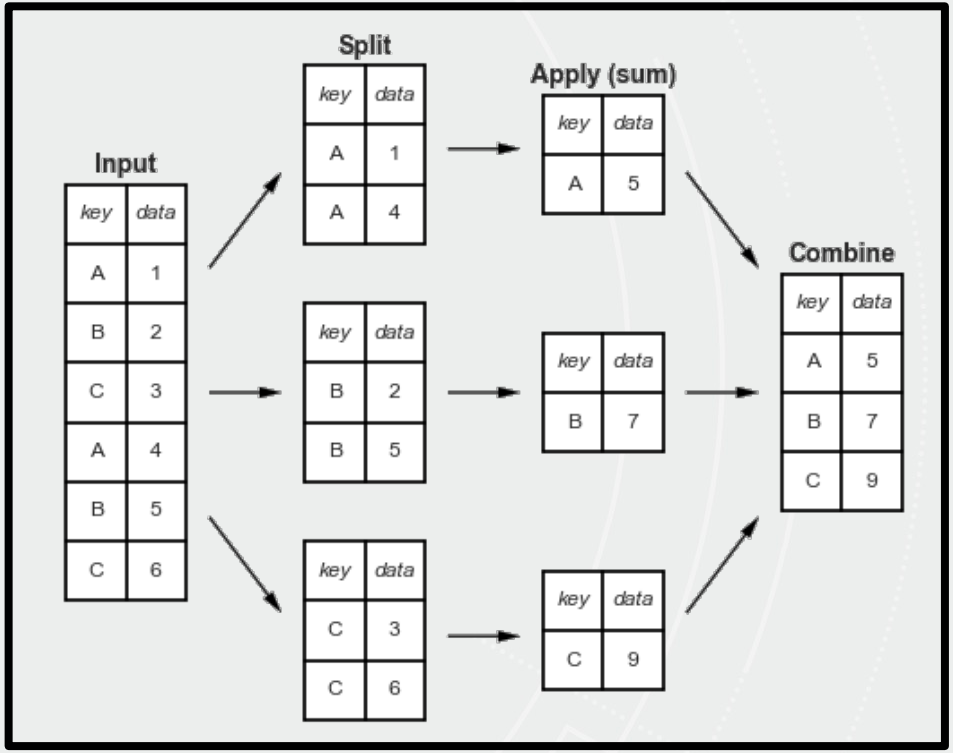


Diagram by Jake VanderPlas.

Second Exercise – Step 1:
Grouping, Pivoting, Resampling and Aggregating

notebooks/am2-data-qaqc.ipynb

Status

Checkpoint

Distributions of Grouped Data (20 mins)

Second Exercise – Step 2:
Statistical Distributions from Regrouped Data

notebooks/am2-data-qaqc.ipynb

Status

Checkpoint

Optional:
Extract These Steps out to a Separate Function

Schedule

AWST	AEST	Agenda	Facilitator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Munging Tabular Data**	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	**Grouping & Reshaping**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	**Explaining Data**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Practice Explaining Your Own Data**	Jeremy
14:45	16:45	**Closeout** – Reflections, Takeaways & Project Update	Tamryn
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Exploring, Documenting & Communicating Data

Develop an Overview of Your Datasets and Explain Some of the Concepts

- The next exercise is about documenting and communicating the features of your dataset in a slightly more structured way
- We'll load in **some of your data**, complete some basic munging, put together a basic graphical overview of your dataset, and add some data documentation
- We'll take the opportunity to run through each person's notebook
- Read the docs if you get stuck or want to find something new!

Exercise

Data Reporting Example (~20mins run-through)

pm2_1-datareport-lithogeochemistry.ipynb

Status

Checkpoint

Note

Template for your Reporting (until break):
pm1-datareport-template.ipynb

Fill in what's relevant and do some exploration of your own.
Feel free to edit, add and remove sections as suits!

Status

Checkpoint

Schedule

AWST	AEST	Agenda	Facilitator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Munging Tabular Data**	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	**Grouping & Reshaping**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	**Explaining Data**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Practice Explaining Your Own Data**	Jeremy
14:45	16:45	**Closeout** – Reflections, Takeaways & Project Update	Tamryn
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Practice Explaining Data (50 mins)

Share

- Run through the key aspects of your data report with the group (5 mins).
- Make notes on dataset features, data issues and snippets of code which could be relevant or useful for you.

Discussion

Did you come up with any interesting ideas, questions, issues and any cool features you've found in the docs?

- Is documenting this process a useful exercise?
- Would your documentation be sufficient to restart after a break? Or for someone else to pick up where you left off?

Case Study (20 mins)

Schedule

AWST	AEST	Agenda	Facilitator
07:30	**09:30**	**Q&A, Issues & Announcements**	
07:45	09:45	**Munging Tabular Data**	Jeremy
09:15	11:15	*Morning Tea*	
09:30	11:30	**Grouping & Reshaping**	Jeremy
11:00	13:00	*Lunch*	
11:45	13:45	**Explaining Data**	Jeremy
13:15	15:15	*Afternoon Tea*	
13:30	15:30	**Practice Explaining Your Own Data**	Jeremy
14:45	16:45	**Closeout** – Reflections, Takeaways & Project Update	Tamryn
14:55	16:55	**Menti**	Tamryn
17:00	17:00	**Close**	

Takeaways & Closeout

Takeaways From Today

- Covered reading data and exploring its properties in pandas
- Covered data types and plotting to visualize relationships
- Reshaping data into new shapes to show other relationships
- Discussed managing data and code once we've done the cleaning process
- Practised telling others about what we've learnt

Next week: Simple Predictions

- Getting into making some models
- Ins and outs of regression
- Choosing between different models
- Dealing with missing data

Capstone Projects

Capstone Projects

Update

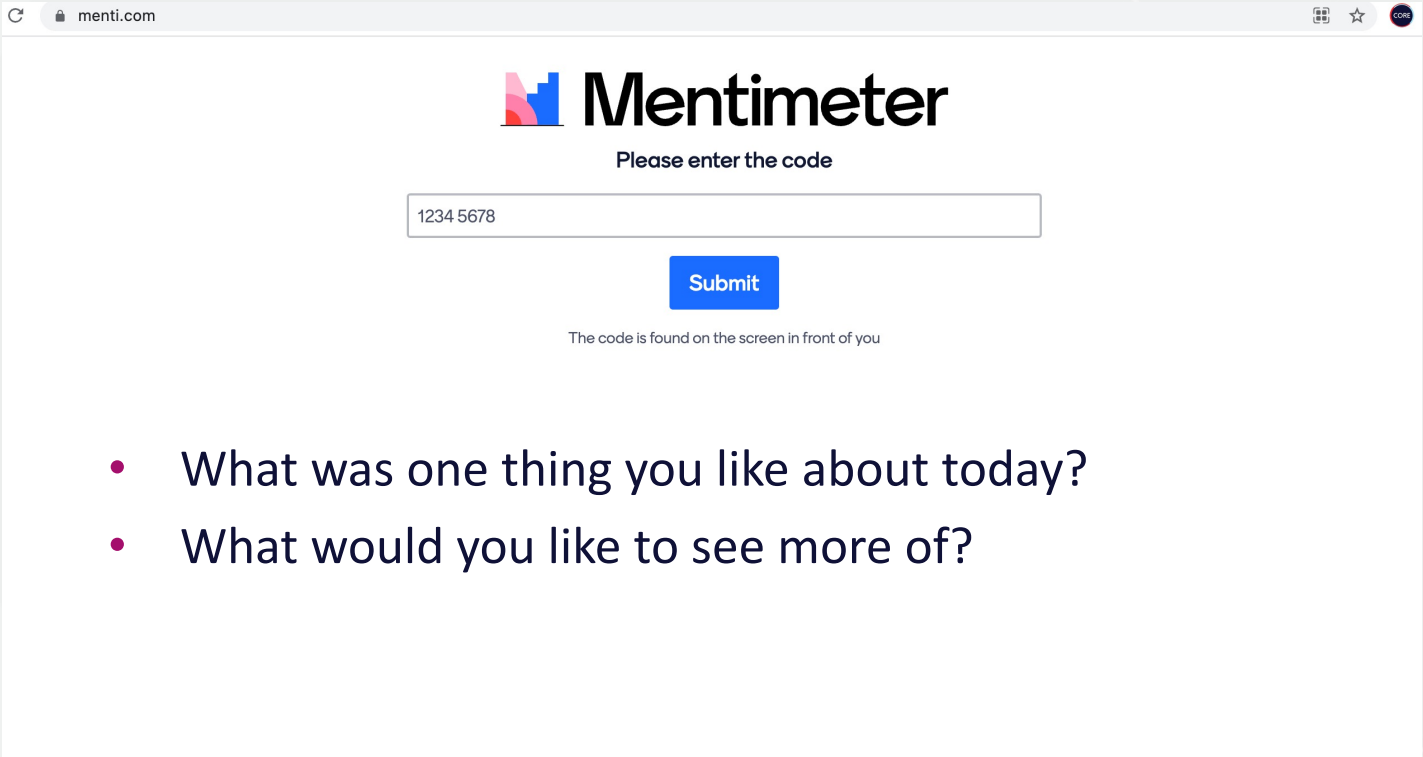
- Your first Project Update (have a go!), how are you scoping your project?

Action

- Milestone #1 → L0 in Momentum
- Update your Working Project Title & Short Project Statement (Problem/Solution/Plan) here.
- Update your Leader

Daily Feedback

Menti



- What was one thing you like about today?
- What would you like to see more of?

CORE Skills

A CORE partnership

COREHUB.COM.AU/SKILLS