Lab session 2

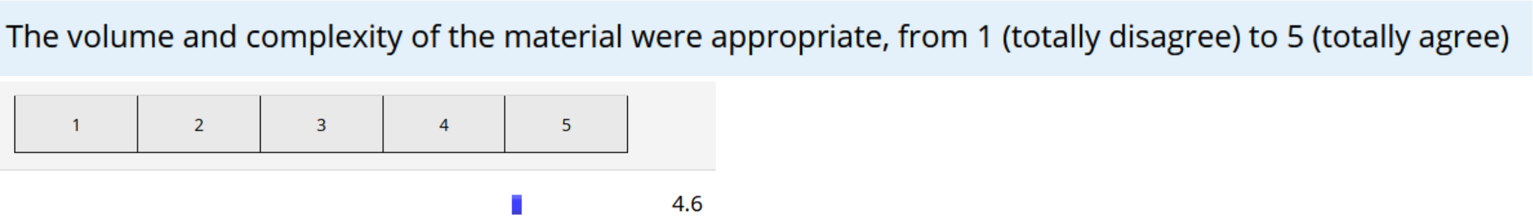
Machine Learning for Behavioral Data (CS-421)

March 03, 2021

Today

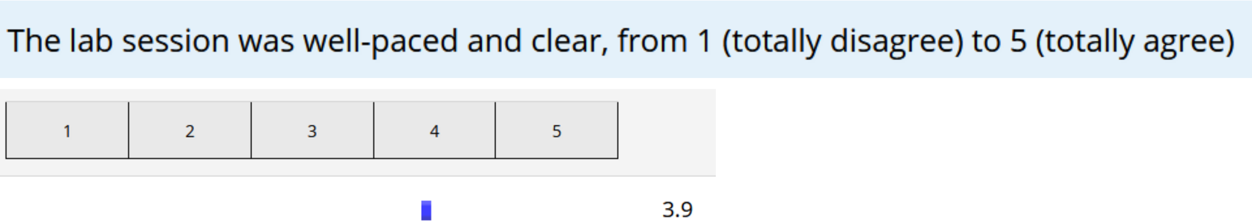
- Debrief last tutorial
- Recap data problems
- Introduction to Pandas
  - SpeakUp and Hands-on
- Data visualization
  - SpeakUp and Hands-on
- Tutorial 2: YouTube trending videos part 2
- Presentation Homework 1

Feedback last tutorial (n=9)



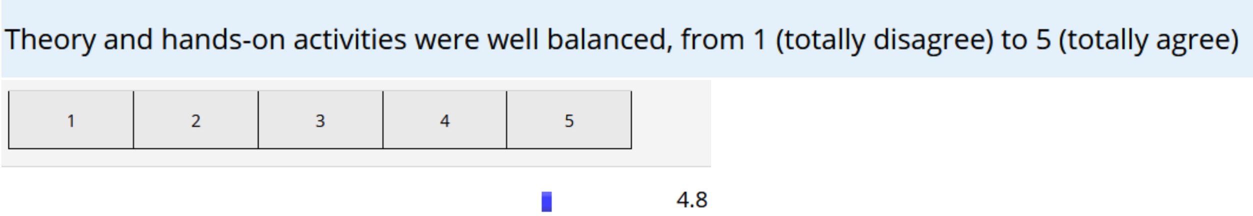
The volume and complexity of the material were appropriate, from 1 (totally disagree) to 5 (totally agree)

4.6



The lab session was well-paced and clear, from 1 (totally disagree) to 5 (totally agree)

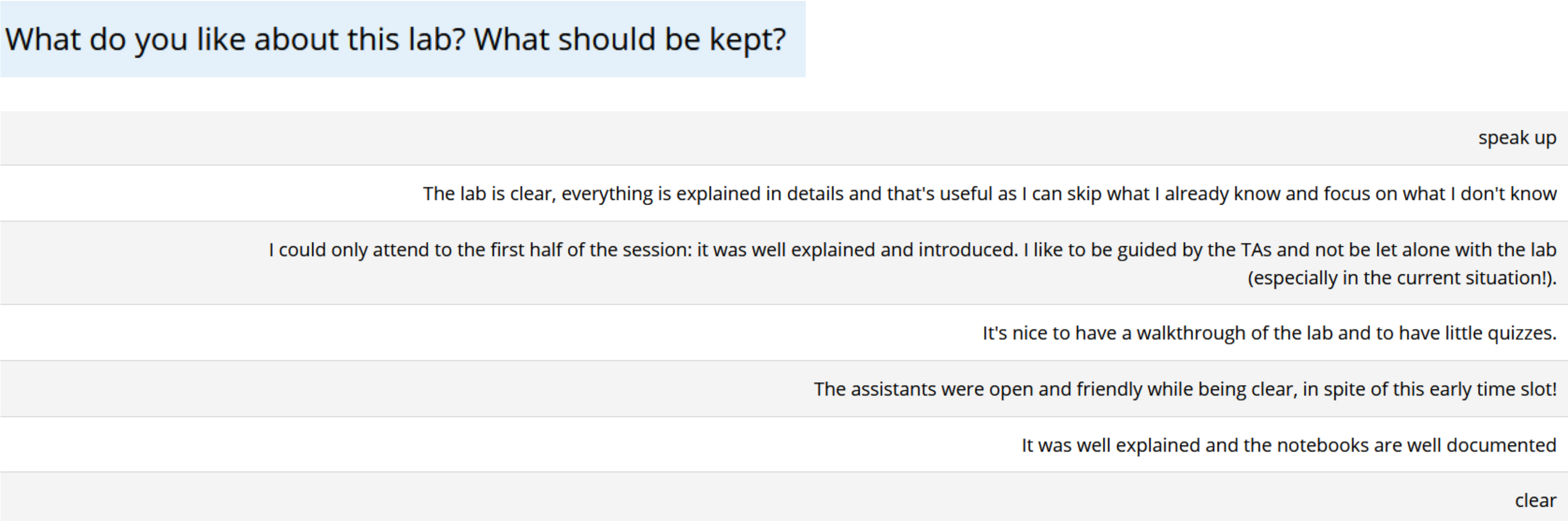
3.9



Theory and hands-on activities were well balanced, from 1 (totally disagree) to 5 (totally agree)

4.8

Feedback last tutorial (n=9)



What do you like about this lab? What should be kept?

speak up

The lab is clear, everything is explained in details and that's useful as I can skip what I already know and focus on what I don't know

I could only attend to the first half of the session: it was well explained and introduced. I like to be guided by the TAs and not be let alone with the lab (especially in the current situation!).

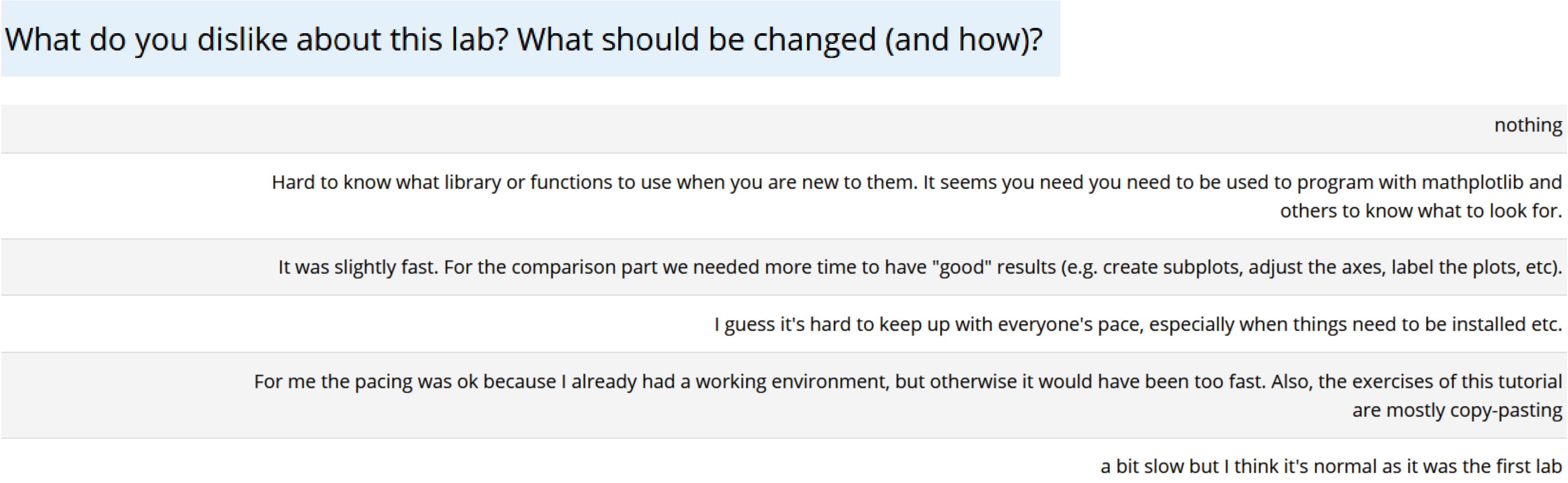
It's nice to have a walkthrough of the lab and to have little quizzes.

The assistants were open and friendly while being clear, in spite of this early time slot!

It was well explained and the notebooks are well documented

clear

Feedback last tutorial (n=9)



What do you dislike about this lab? What should be changed (and how)?

nothing

Hard to know what library or functions to use when you are new to them. It seems you need you need to be used to program with mathplotlib and others to know what to look for.

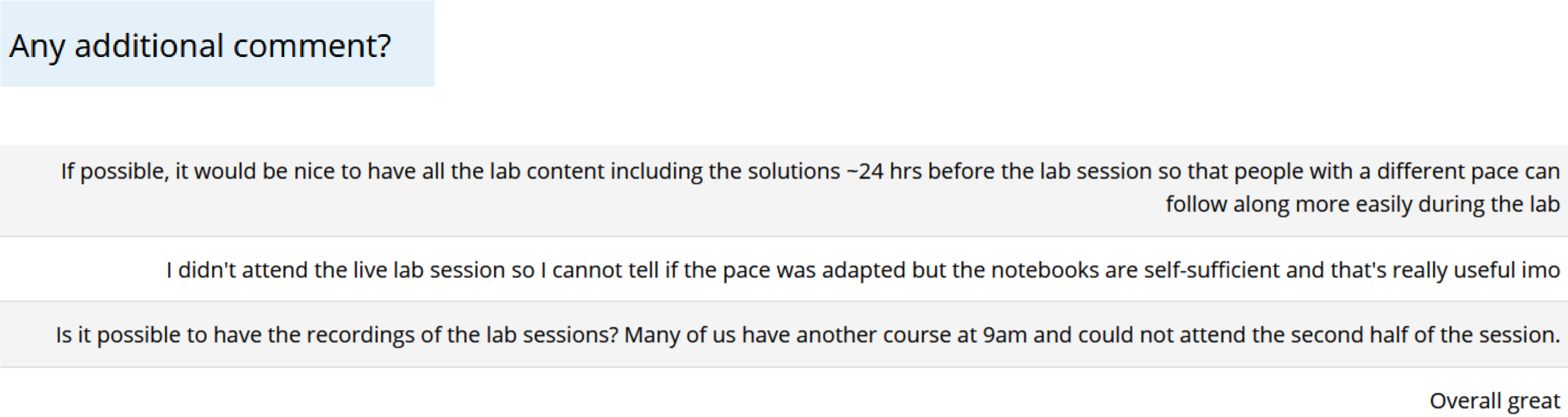
It was slightly fast. For the comparison part we needed more time to have "good" results (e.g. create subplots, adjust the axes, label the plots, etc).

I guess it's hard to keep up with everyone's pace, especially when things need to be installed etc.

For me the pacing was ok because I already had a working environment, but otherwise it would have been too fast. Also, the exercises of this tutorial are mostly copy-pasting

a bit slow but I think it's normal as it was the first lab

Feedback last tutorial (n=9)



Any additional comment?

If possible, it would be nice to have all the lab content including the solutions ~24 hrs before the lab session so that people with a different pace can follow along more easily during the lab

I didn't attend the live lab session so I cannot tell if the pace was adapted but the notebooks are self-sufficient and that's really useful imo

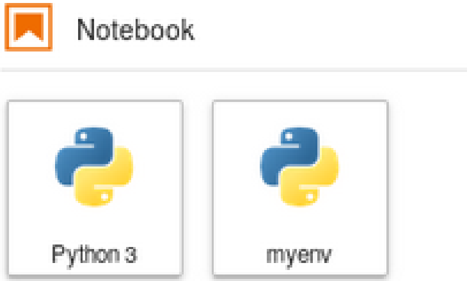
Is it possible to have the recordings of the lab sessions? Many of us have another course at 9am and could not attend the second half of the session.

Overall great

Addendum Tutorial 1.2

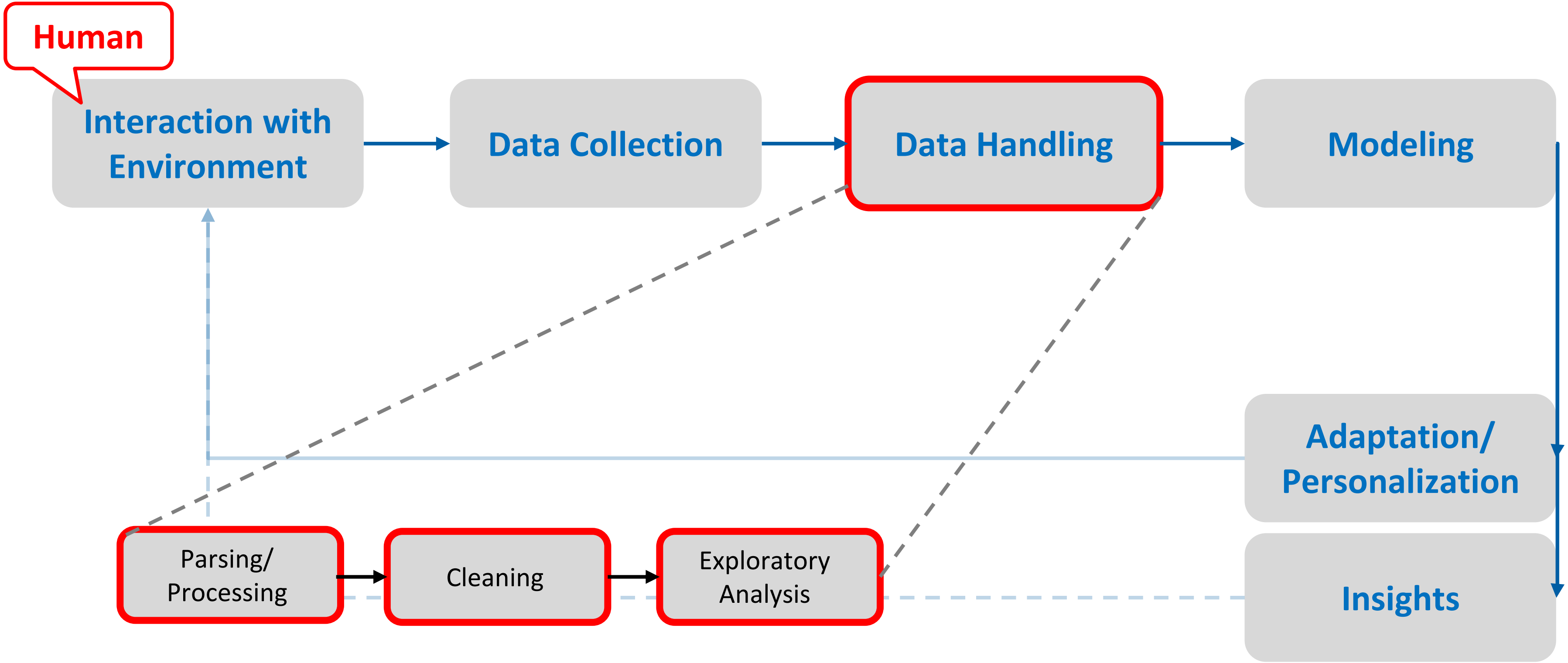
- Virtual environment:
  - https://janakiev.com/blog/jupyter-virtual-envs/
  - Create virtual environment: `python -m venv myenv`
  - Activate virtual environment: `source myenv/bin/activate`
  - add to Jupyter (deactivate virtual environment first)

```
python -m ipykernel install --user --name=myenv
```

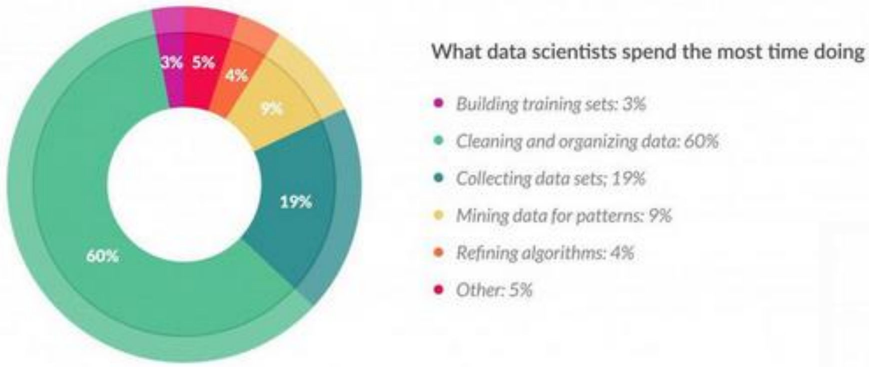


Questions?

What is ML for Behavioral Data?

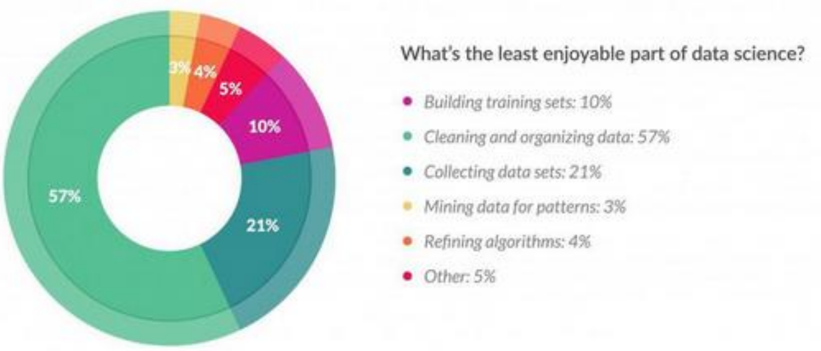


A survey of data scientists



What data scientists spend the most time doing

- ● Building training sets: 3%
- ● Cleaning and organizing data: 60%
- ● Collecting data sets; 19%
- ● Mining data for patterns: 9%
- ● Refining algorithms: 4%
- ● Other: 5%



What's the least enjoyable part of data science?

- ● Building training sets: 10%
- ● Cleaning and organizing data: 57%
- ● Collecting data sets: 21%
- ● Mining data for patterns: 3%
- ● Refining algorithms: 4%
- ● Other: 5%

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=1ca6c58d6f63

Data Problems

- Incorrect data
- Duplicates
- Inconsistent data
- Missing data
- Outliers

Data Problems

- **Incorrect data**
- Duplicates
- Inconsistent data
- Missing data
- Outliers

User	Hours spent on platform **per day**
1	3
2	2
3	3
4	**9999**
5	4
6	3

Data Problems

- Incorrect data
- **Duplicates**
- Inconsistent data
- Missing data
- Outliers

Username	Email	Number of visits
mmarras	mirko.marras@epfl.ch	4
mirko	m.marras@gmail.com	3
mirko marras	mirko.marras@epfl.com	2
chris g.	christian.giang@epfl.ch	6

Data Problems

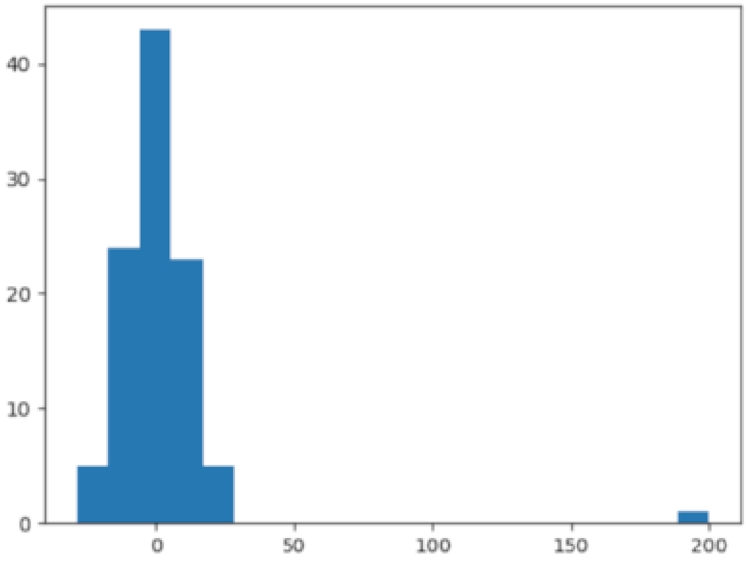
- Incorrect data
- Duplicates
- **Inconsistent data**
- Missing data
- Outliers

User	Auto-reported duration
1	10
2	Half an hour
3	From 15-30 min
4	2hrs
5	Less than a day

Nonstandard units

Data Problems

- Incorrect data
- Duplicates
- Inconsistent data
- Missing data
- **Outliers**



Data Problems

- Incorrect data
- Duplicates
- Inconsistent data
- **Missing data**
- Outliers

User	Auto-reported duration	Hours spent on platform
1	**nan**	3
2	Half an hour	
3	**None**	**na**
4	**N/A**	**9999**
5	**NaN**	-

Missing Data | How does it look like?

User	Auto-reported duration	Hours spent on platform
1	**nan**	$
2	X	
3	**None**	**na**
4	**N/A**	**9999**
5	**NaN**	-

Some tools might not identify all the different forms of missing values.

Can take multiple forms

Data inspection is crucial.

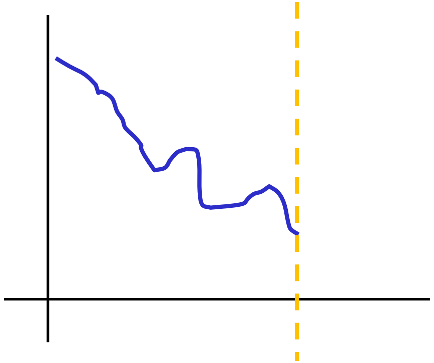
Missing Data | Best practices

Feature	Mean	Percentage of missing values
Auto-reported duration	30	**3%**
Hours spent on platform	4.5	**40%**

1. Report the percentage of missing data per feature or observation.

2. Try to understand why the data is missing?

Attrition, non-response

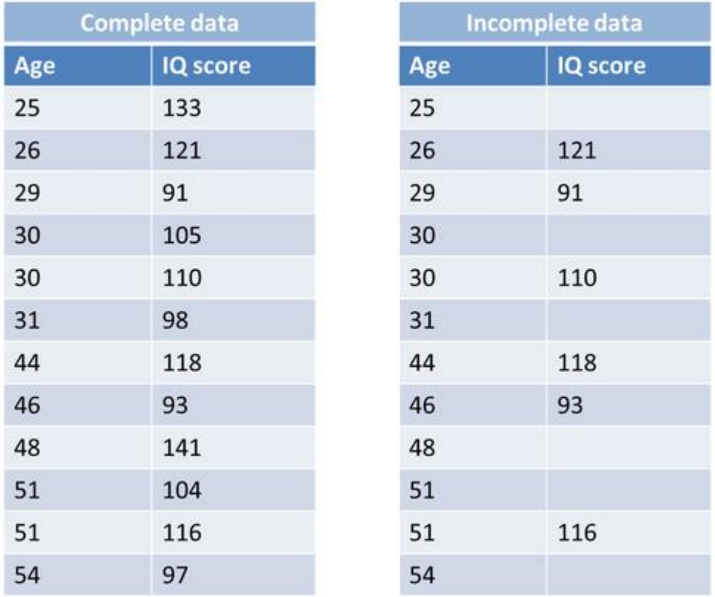


Non-stochastic and stochastic imputation methods

Listwise deletion

3. Determine the most appropriate method for handling missing data

Patterns of missing data



Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

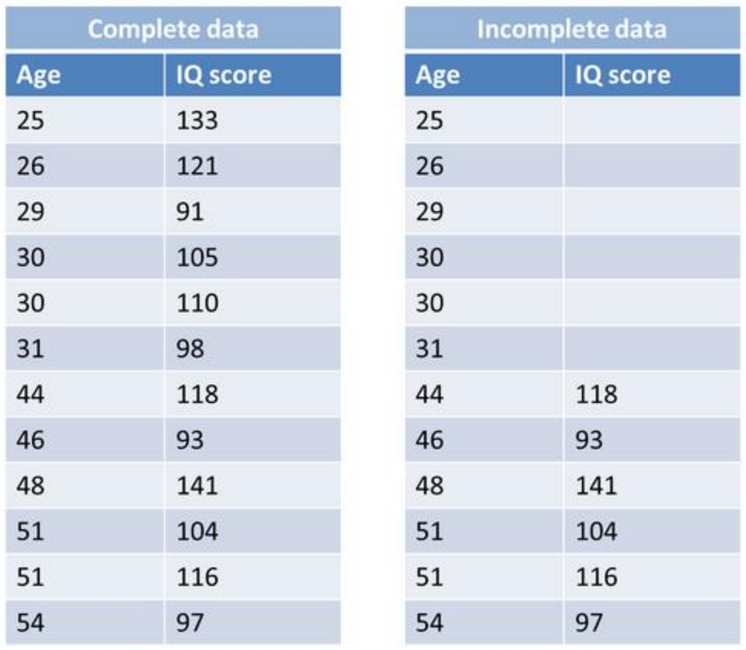
Example of MCAR data

https://www.iriseekhout.com/missing-data/missing-data-mechanisms/mcar/

MCAR:

- Missing completely at random
- Test by separating and examining missing and complete cases

Patterns of missing data



Complete data		Incomplete data	
Age	IQ score	Age	IQ score
25	133	25	
26	121	26	
29	91	29	
30	105	30	
30	110	30	
31	98	31	
44	118	44	118
46	93	46	93
48	141	48	141
51	104	51	104
51	116	51	116
54	97	54	97

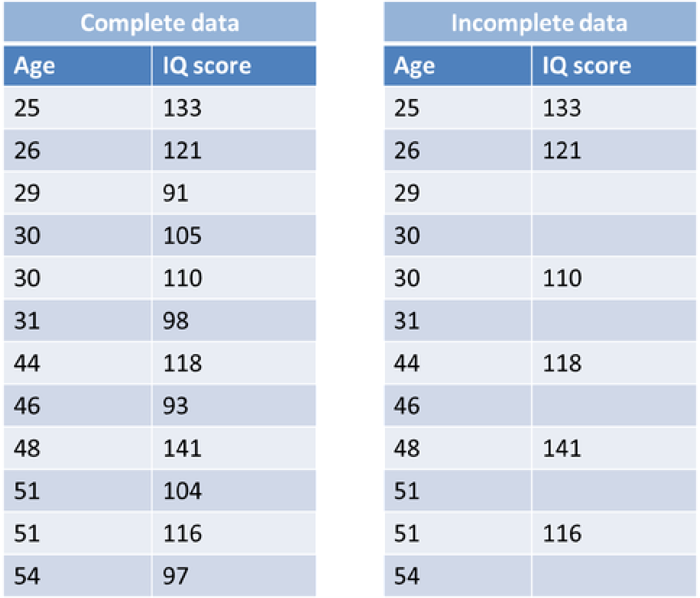
Example of MAR data

https://www.iriseekhout.com/missing-data/missing-data-mechanisms/mar/

MAR:

- Missing at random
- Missing related to other measured variable but not to variable with missing values itself

Patterns of missing data



Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	133
26	121
29	
30	
30	110
31	
44	118
46	
48	141
51	
51	116
54	

Example of MNAR data

https://www.iriseekhout.com/missing-data/missing-data-mechanisms/mnar/

MNAR (or NMAR):

- Missing not at random
- Missing related to variable with missing values itself

Patterns of missing data

- How to know which assumption to make?
- Know your data!
  - Knowledge about the data collection
  - Knowledge about the domain
- Statistical tests
- “Common sense”

Handling missing data

- There is not ONE correct method
  - It depends on your data
- Example methods from Schlomer and Bauman 2010[1]:

Deletion

Listwise deletion

Pairwise deletion

Non-stochastic imputation

Mean substitution

Regression substitution

Pattern-matching imputation

Stochastic imputation

Stochastic regression

Expectation maximization

Multiple imputation

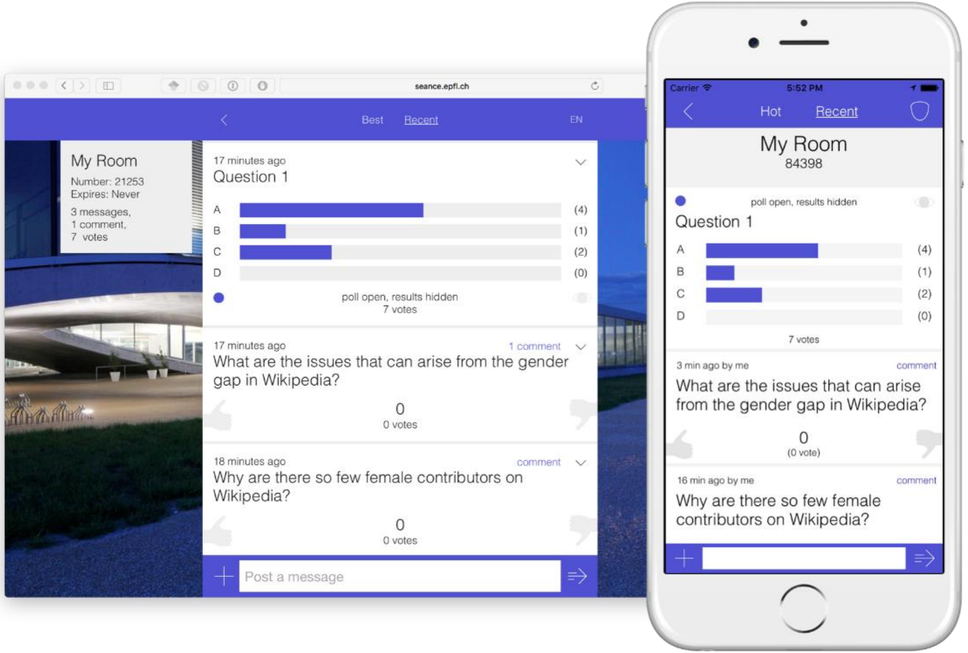
Full information maximum likelihood

[1] https://www.researchgate.net/publication/49660163_Best_Practices_for_Missing_Data_Management_in_Counseling_Psychology

SpeakUp

- **Android / iOS:**
  http://speakup.info/
- **Web App:**
  https://web.speakup.info/

- Room number: ???



My Room
Number: 21253
Expires: Never
3 messages,
1 comment,
7 votes

17 minutes ago
Question 1

A		(4)
B		(1)
C		(2)
D		(0)

poll open, results hidden
7 votes

17 minutes ago — 1 comment
What are the issues that can arise from the gender gap in Wikipedia?
0
0 votes

18 minutes ago — comment
Why are there so few female contributors on Wikipedia?
0
0 votes

Post a message

My Room
84398

poll open, results hidden
Question 1

A		(4)
B		(1)
C		(2)
D		(0)

7 votes

3 min ago by me — comment
What are the issues that can arise from the gender gap in Wikipedia?
0
(0 vote)

16 min ago by me — comment
Why are there so few female contributors on Wikipedia?

Pandas

SpeakUp: How do you feel about Pandas?

A: I think it’s sad there are only around 2000 of them living in the wild.

B: I’m **not confident at all** about using Pandas.

C: I’m **slightly confident** about using Pandas.

D: I’m **fairly confident** about using Pandas.

E: I’m **very confident** about using Pandas.

Quiz

Q1: How many different data structures does Pandas provide?

A: 1

B: 2 ✓

C: 3

D: 4

Quiz

Q2: Which command allows you to inspect some row entries of a dataframe `df`?

A: `df.describe()`

B: `df.info()`

[✓ C: `df.tail()`]

D: `df.show()`

Quiz

Q3: Which command returns the first two rows of a Pandas dataframe `df`?

A: `df.iloc[2]`

B: `df.head()`

C: `df.loc[0:1]` ✓

D: `df.loc[0,1]`

Quiz

Q4: How would you remove rows of a dataframe `df` containing some empty values ?

A: `df.dropna(axis=0, how=‘all’)`

B: `df.dropna(axis=0, how=‘any’)` ✓

C: `df.dropna(axis=1, how=‘any’)`

D: `df.dropna(axis=1, how=‘all’)`

Quiz

Q5: How would you find duplicates in a dataframe `df`?

A: `df.duplicates()`

B: `df.duplicated()` ✓

C: `df.duplicate()`

D: `df.dup()`

Quiz

Q6: How would you replace empty values in a dataframe `df` with the value `1234`?

A: `df.fillna(1234)` ✓

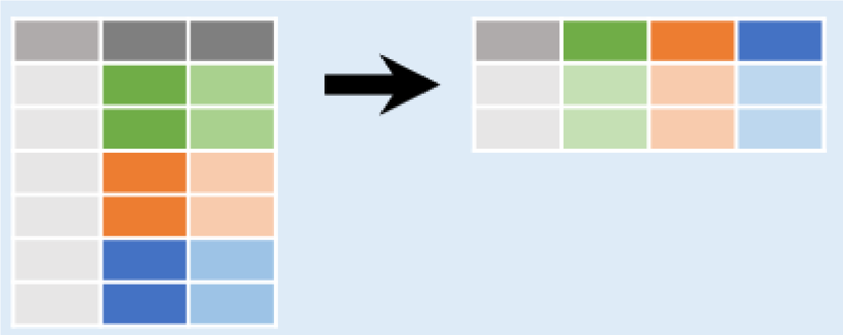
B: `df.fillna(1234, np.nan)`

C: `df.replacena(1234)`

D: `df.replace(1234, np.nan)`

Quiz

Q7: Which Pandas function performs the following operation on the dataframe `df`?



A: `pd.melt(df)`

B: `pd.concat(columns=‘var’)`

C: `df.groupby(by=“col”)`

D: `df.pivot(columns=‘var’, values=‘val’)` ✓

Pandas resources

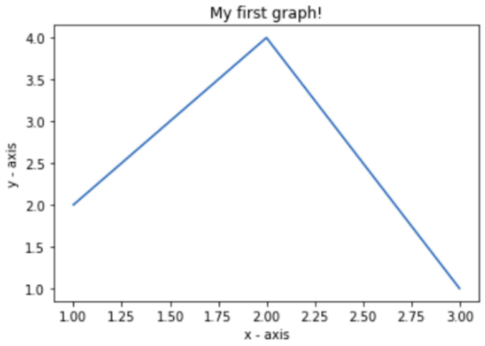
- Choose from below according to your level:
  - https://www.w3schools.com/python/pandas/default.asp
    - 14 beginner-level Pandas tutorial pages
    - Self-evaluation test with 25 questions
  - https://www.javatpoint.com/python-pandas
    - More exhaustive collection of tutorials
    - Check out the “Pandas Interview Questions”
  - https://pandas.pydata.org/Pandas_Cheat_Sheet.pdf
    - Pandas cheat sheet on two A4 pages

Data visualization

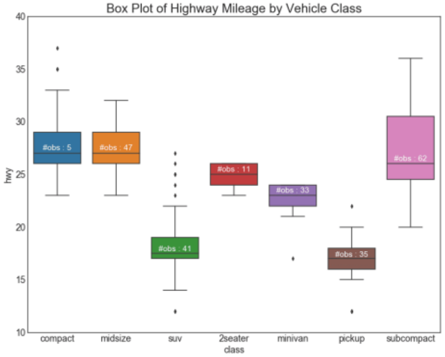
SpeakUp:

Which figure represents your Python data visualization skills?

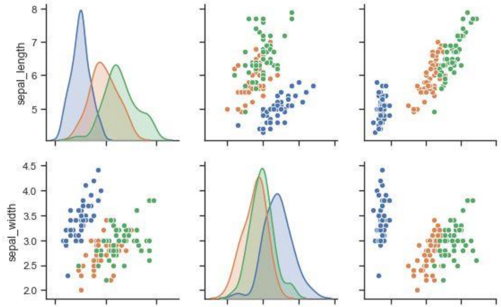
A



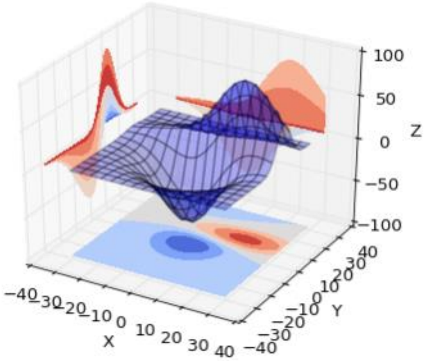
B



C



D



Matplotlib

SpeakUp: How do you feel about Matplotlib?

A: I’m **not confident at all** about using Matplotlib.

B: I’m **slightly confident** about using Matplotlib.

C: I’m **fairly confident** about using Matplotlib.

D: I’m **very confident** about using Matplotlib.

https://matplotlib.org/stable/index.html

Seaborn

SpeakUp: How do you feel about Seaborn?

A: I’m **not confident at all** about using Seaborn.

B: I’m **slightly confident** about using Seaborn.

C: I’m **fairly confident** about using Seaborn.

D: I’m **very confident** about using Seaborn.

http://seaborn.pydata.org/index.html

Bokeh

SpeakUp: How do you feel about Bokeh?

A: I’m **not confident at all** about using Bokeh.

B: I’m **slightly confident** about using Bokeh.

C: I’m **fairly confident** about using Bokeh.

D: I’m **very confident** about using Bokeh.

bokeh

https://bokeh.org/

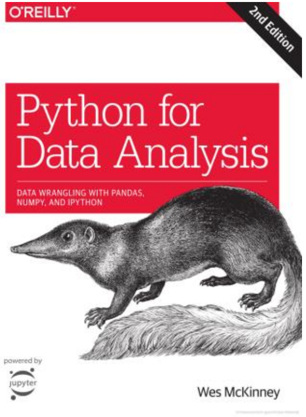
Data visualization resources

- **Matplotlib**
  - Gallery of plots and tutorial: https://matplotlib.org/stable/index.html
- **Seaborn**
  - Gallery of plots and tutorial: http://seaborn.pydata.org/index.html
- **Bokeh**
  - Gallery of plots and tutorial: https://docs.bokeh.org/en/latest/index.html
- **Cheat sheets**
  - Matplotlib: https://blog.finxter.com/best-matplotlib-cheat-sheet/
  - Seaborn: https://www.datacamp.com/community/blog/seaborn-cheat-sheet-python
  - Bokeh: https://www.datacamp.com/community/blog/bokeh-cheat-sheet-python

Tutorial 2

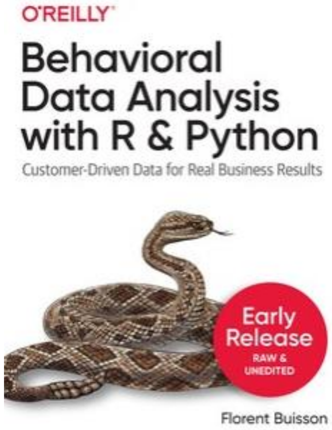
- YouTube trending videos datasets for CA
- **Task**: Pull Tutorial 2 from GitHub and work through the notebook step-by-step. After each step, we will discuss all together.

Further readings



McKinney, W. (2017). **Python for Data Analysis: Data Wrangling with Pandas, Numpy, and IPython**. 2/ed. O'Reilly Media.

Buisson, F. (2021). **Behavioral Data Analysis with R and Python**. O'Reilly Media.

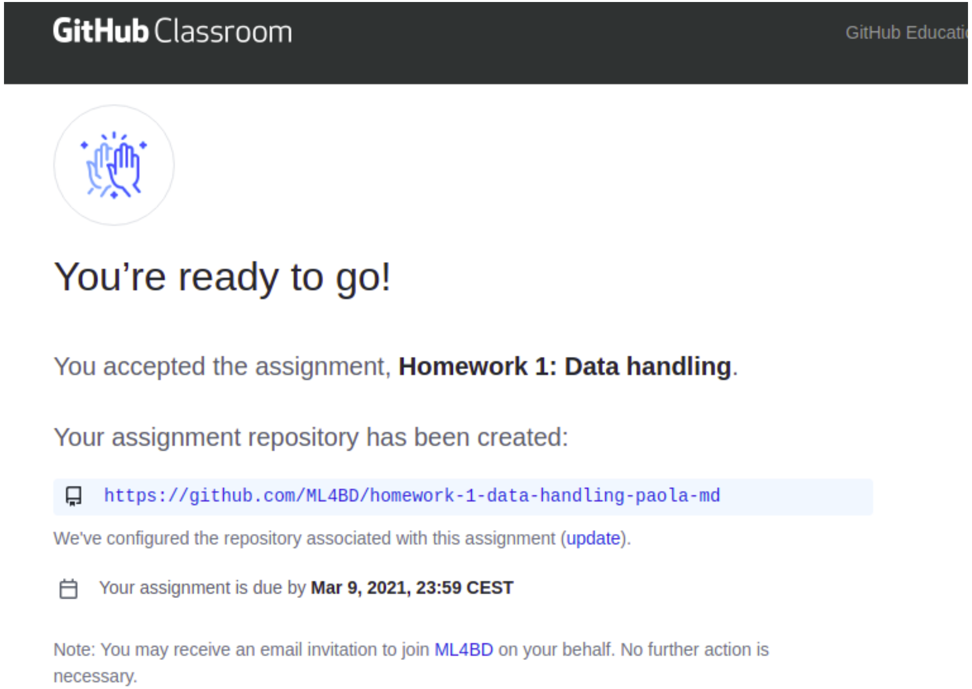

O'REILLY
Behavioral Data Analysis with R & Python
Customer-Driven Data for Real Business Results
Early Release
RAW & UNEDITED
Florent Buisson

Homework 1

Link for homework: https://classroom.github.com/a/QM_Oahl8

Complete the exercises in the .ipynb notebook, commit your work, and **push the changes to github**. The TAs will pull the completed assignment after the deadline Mar 09, 2021 23:59 CET (any changes after the deadline will not be considered)

If you have any questions, feel free to use the Q&A forum in Moodle!



You're ready to go!

You accepted the assignment, **Homework 1: Data handling**.

Your assignment repository has been created:

https://github.com/ML4BD/homework-1-data-handling-paola-md

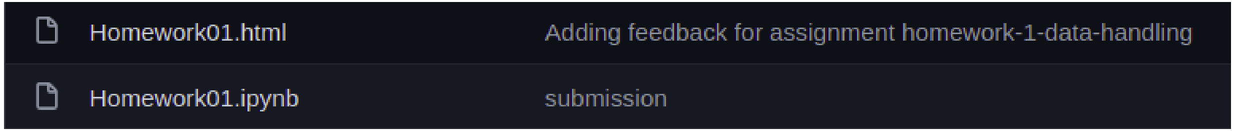
We've configured the repository associated with this assignment (update).

Your assignment is due by **Mar 9, 2021, 23:59 CEST**

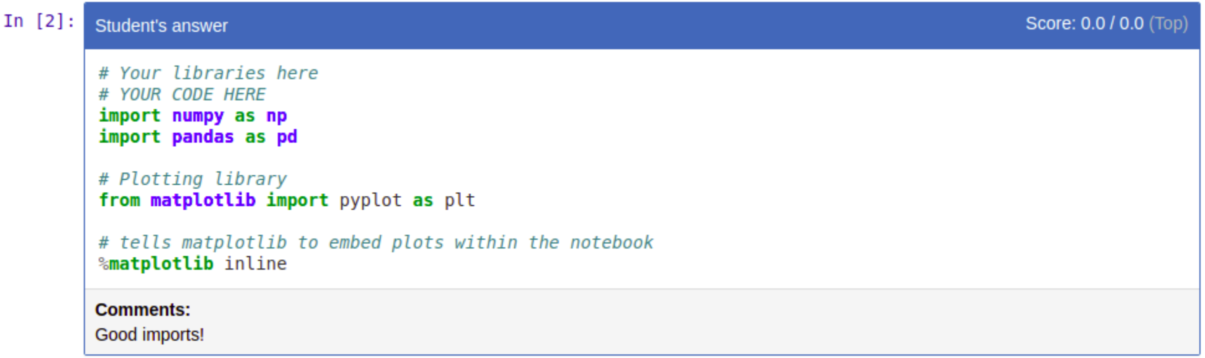
Note: You may receive an email invitation to join ML4BD on your behalf. No further action is necessary.

Homework Grade

We will push the graded homework with feedback to each repo.
You will see comments and number of points awarded.



Homework01.html	Adding feedback for assignment homework-1-data-handling
Homework01.ipynb	submission



In [2]:

Student's answer

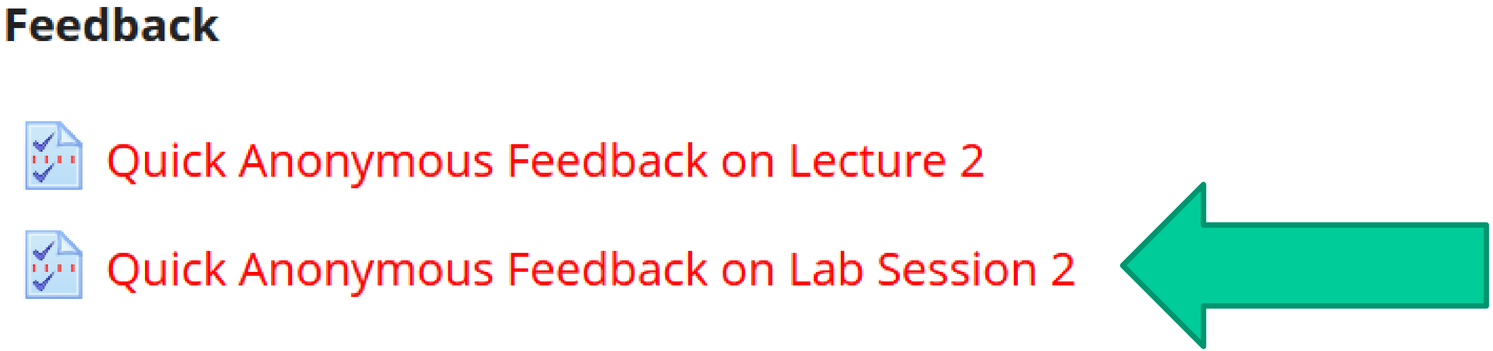
Score: 0.0 / 0.0 (Top)

```
# Your libraries here
# YOUR CODE HERE
import numpy as np
import pandas as pd

# Plotting library
from matplotlib import pyplot as plt

# tells matplotlib to embed plots within the notebook
%matplotlib inline
```

Comments:
Good imports!

Feedback

Please give feedback on this lab session:
https://moodle.epfl.ch/mod/questionnaire/view.php?id=1134450



Feedback

Quick Anonymous Feedback on Lecture 2

Quick Anonymous Feedback on Lab Session 2

Questions?