Lab session 4

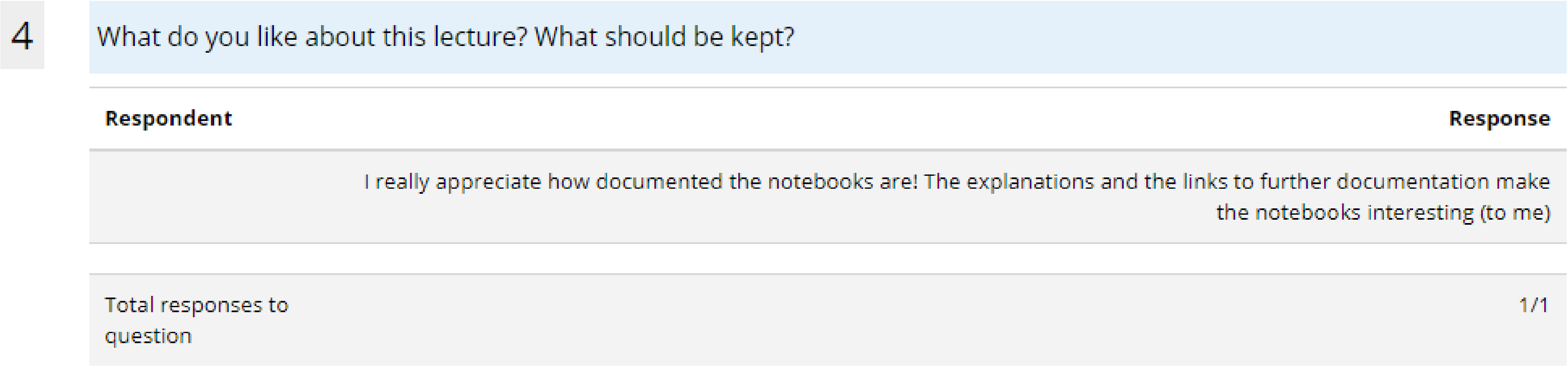
Machine Learning for Behavioral Data (CS-421)

March 17, 2021

Today

- **08:15 08:45 Part I Debriefing Previous Week**
  - Feedback from tutorial 3 and homework 1
  - More extensive debriefing on homework 1
- **08:45 09:05 Part II Tutorial on Model Evaluation**
- **09:05 09:10 SHORT BREAK**
- **09:10 09:30 Part III Tutorial on Model Selection**
- **09:30 10:00 Part IV Introduction to Next Steps**
  - Class project presentation
  - Introduction to homework 3

Feedback tutorial 3 (n=1)



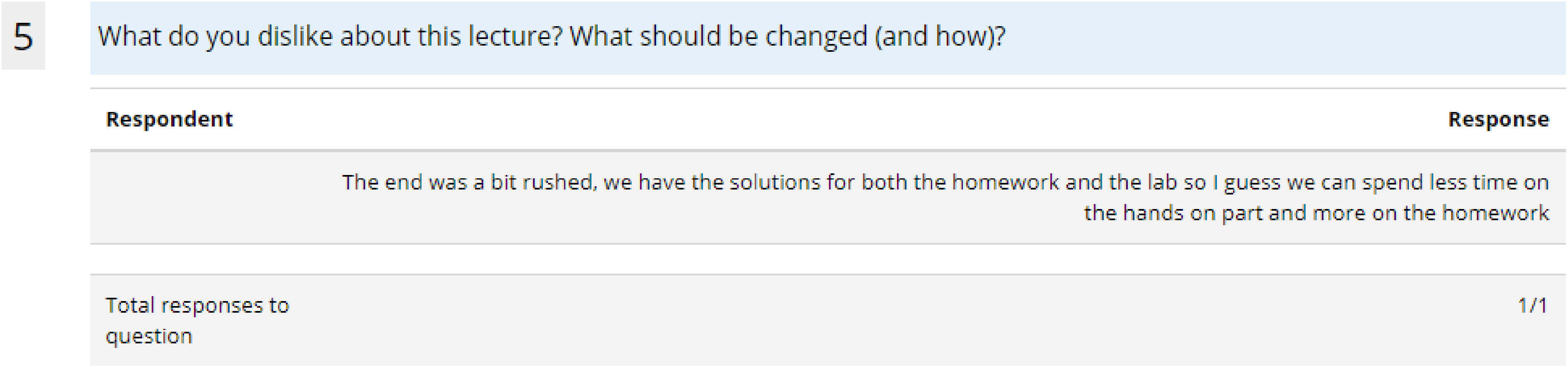
4 What do you like about this lecture? What should be kept?

Respondent	Response
	I really appreciate how documented the notebooks are! The explanations and the links to further documentation make the notebooks interesting (to me)
Total responses to question	1/1

Other points from previous feedback forms:

- Hard to know what library or function to use and when. Provide supporting information.
- Provide a bit more time to complete exercise cells with nice results.
- Control the pace of the tutorial and provide more non-copy-and-pasting exercises.

Feedback tutorial 3 (n=1)



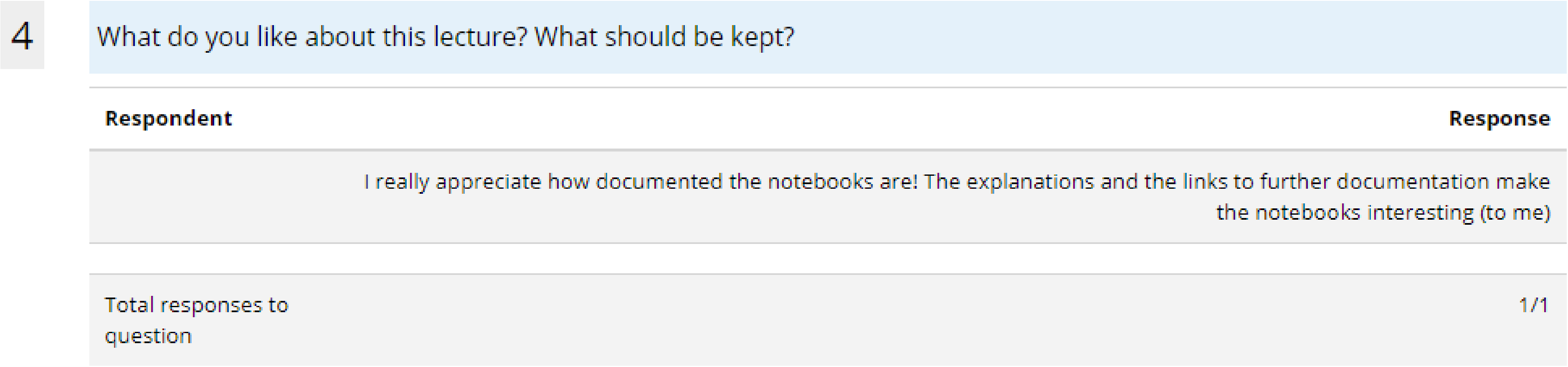
5 What do you dislike about this lecture? What should be changed (and how)?

Respondent	Response
	The end was a bit rushed, we have the solutions for both the homework and the lab so I guess we can spend less time on the hands on part and more on the homework
Total responses to question	1/1

Other points from previous feedback forms:

- Time of uploading for the material.
- Investigating the possibility of recording lab sessions.

Feedback tutorial 3 (n=1)



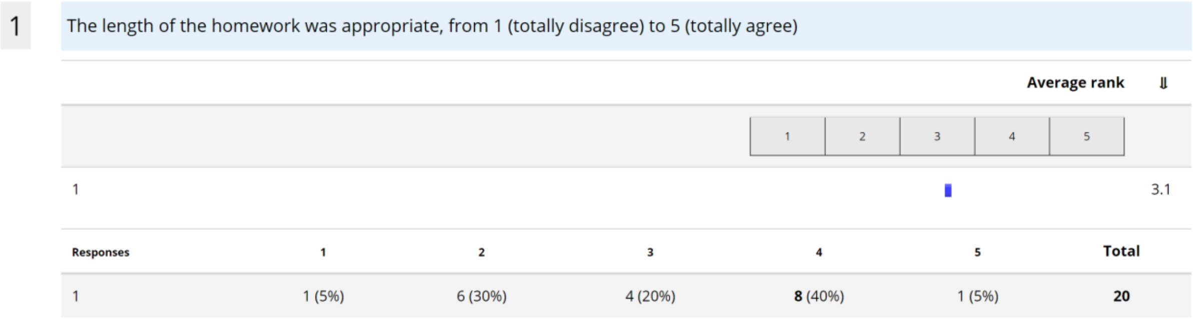
4 What do you like about this lecture? What should be kept?

Respondent	Response
	I really appreciate how documented the notebooks are! The explanations and the links to further documentation make the notebooks interesting (to me)
Total responses to question	1/1

Other points from previous feedback forms:

- Hard to know what library or function to use and when. Provide supporting information.
- Provide a bit more time to complete exercise cells with nice results.
- Control the pace of the tutorial and provide more non-copy-and-pasting exercises.

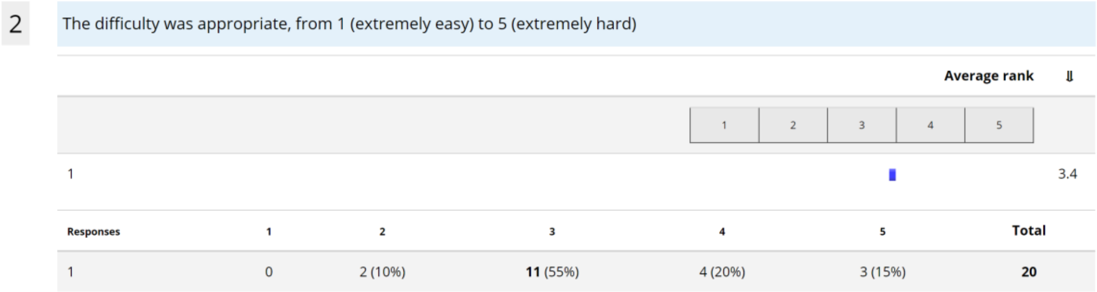
Feedback homework 1 (n=20)



1 The length of the homework was appropriate, from 1 (totally disagree) to 5 (totally agree)

	Average rank ⇓
1	3.1

Responses	1	2	3	4	5	Total
1	1 (5%)	6 (30%)	4 (20%)	**8** (40%)	1 (5%)	**20**

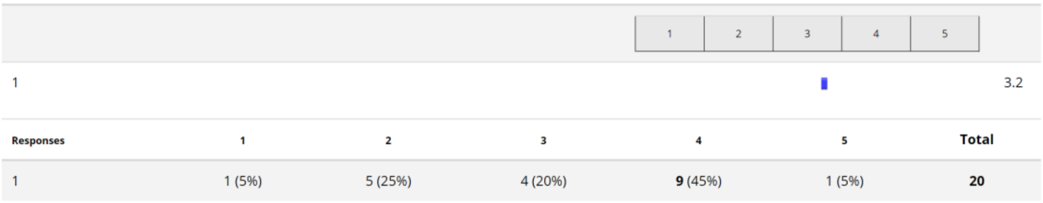


2 The difficulty was appropriate, from 1 (extremely easy) to 5 (extremely hard)

	Average rank ⇓
1	3.4

Responses	1	2	3	4	5	Total
1	0	2 (10%)	**11** (55%)	4 (20%)	3 (15%)	**20**

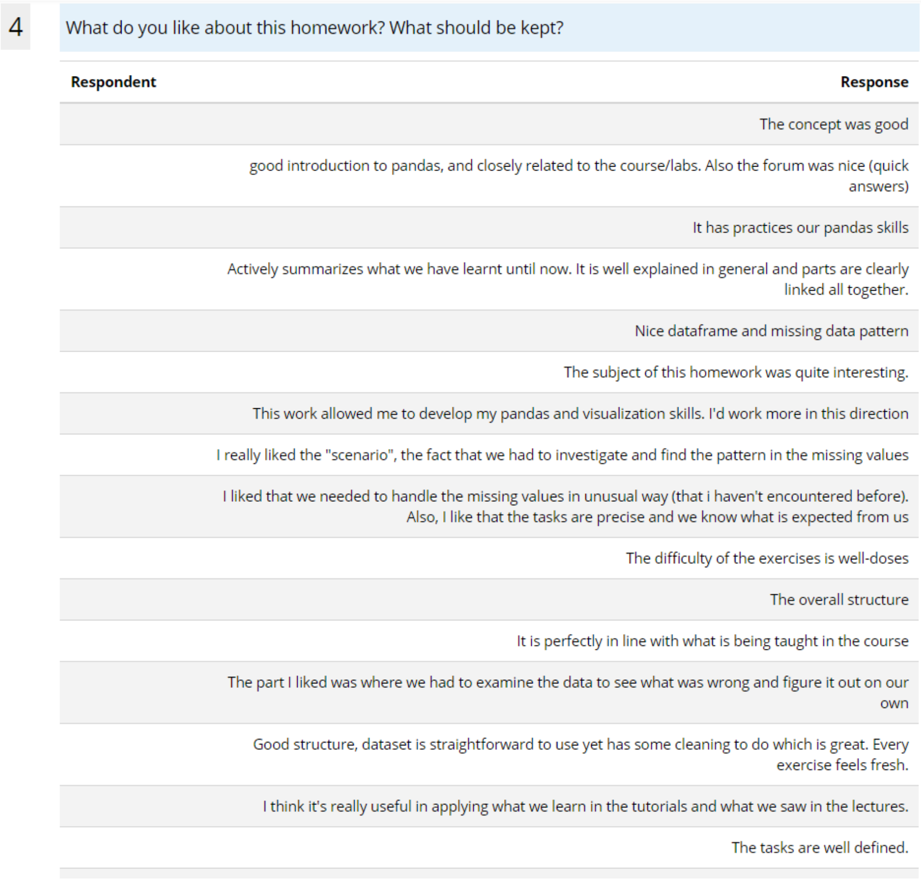
3 The instructions were clear, from 1 (totally disagree) to 5 (totally agree)



1	3.2

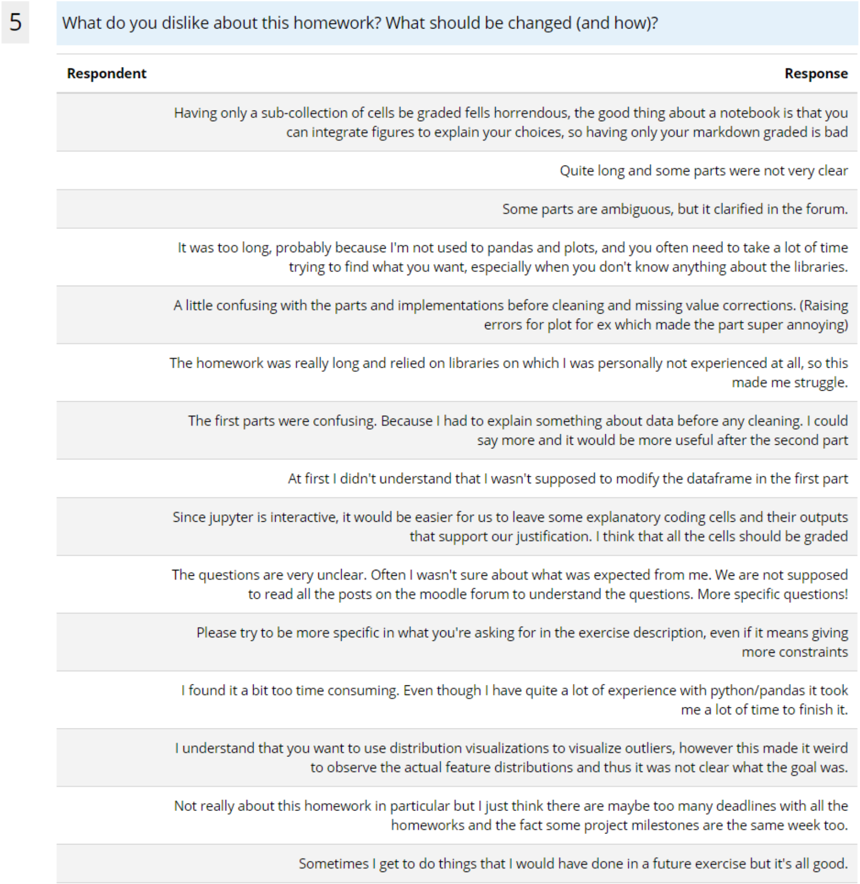
Responses	1	2	3	4	5	Total
1	1 (5%)	5 (25%)	4 (20%)	**9** (45%)	1 (5%)	**20**

Feedback homework 1 (n=20)



4 What do you like about this homework? What should be kept?

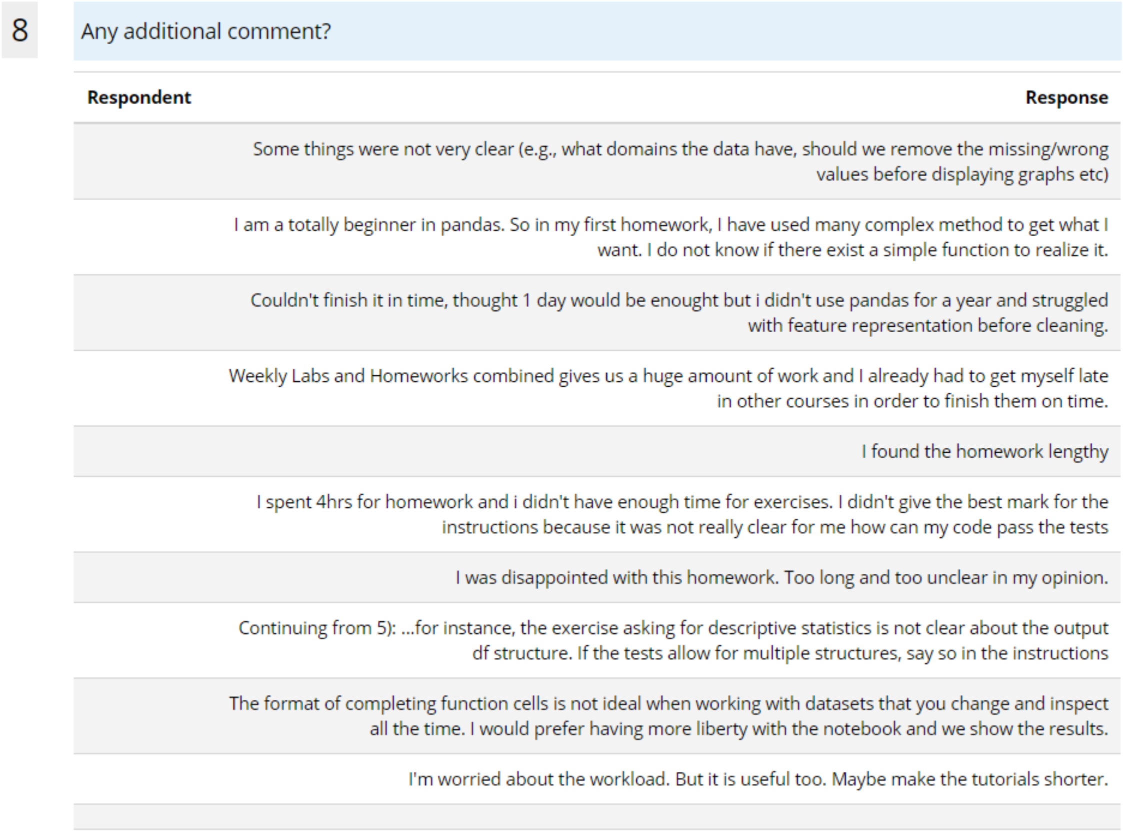
Respondent	Response
	The concept was good
	good introduction to pandas, and closely related to the course/labs. Also the forum was nice (quick answers)
	It has practices our pandas skills
	Actively summarizes what we have learnt until now. It is well explained in general and parts are clearly linked all together.
	Nice dataframe and missing data pattern
	The subject of this homework was quite interesting.
	This work allowed me to develop my pandas and visualization skills. I'd work more in this direction
	I really liked the "scenario", the fact that we had to investigate and find the pattern in the missing values
	I liked that we needed to handle the missing values in unusual way (that i haven't encountered before). Also, I like that the tasks are precise and we know what is expected from us
	The difficulty of the exercises is well-doses
	The overall structure
	It is perfectly in line with what is being taught in the course
	The part I liked was where we had to examine the data to see what was wrong and figure it out on our own
	Good structure, dataset is straightforward to use yet has some cleaning to do which is great. Every exercise feels fresh.
	I think it's really useful in applying what we learn in the tutorials and what we saw in the lectures.
	The tasks are well defined.



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

Respondent	Response
	Having only a sub-collection of cells be graded fells horrendous, the good thing about a notebook is that you can integrate figures to explain your choices, so having only your markdown graded is bad
	Quite long and some parts were not very clear
	Some parts are ambiguous, but it clarified in the forum.
	It was too long, probably because I'm not used to pandas and plots, and you often need to take a lot of time trying to find what you want, especially when you don't know anything about the libraries.
	A little confusing with the parts and implementations before cleaning and missing value corrections. (Raising errors for plot for ex which made the part super annoying)
	The homework was really long and relied on libraries on which I was personally not experienced at all, so this made me struggle.
	The first parts were confusing. Because I had to explain something about data before any cleaning. I could say more and it would be more useful after the second part
	At first I didn't understand that I wasn't supposed to modify the dataframe in the first part
	Since jupyter is interactive, it would be easier for us to leave some explanatory coding cells and their outputs that support our justification. I think that all the cells should be graded
	The questions are very unclear. Often I wasn't sure about what was expected from me. We are not supposed to read all the posts on the moodle forum to understand the questions. More specific questions!
	Please try to be more specific in what you're asking for in the exercise description, even if it means giving more constraints
	I found it a bit too time consuming. Even though I have quite a lot of experience with python/pandas it took me a lot of time to finish it.
	I understand that you want to use distribution visualizations to visualize outliers, however this made it weird to observe the actual feature distributions and thus it was not clear what the goal was.
	Not really about this homework in particular but I just think there are maybe too many deadlines with all the homeworks and the fact some project milestones are the same week too.
	Sometimes I get to do things that I would have done in a future exercise but it's all good.

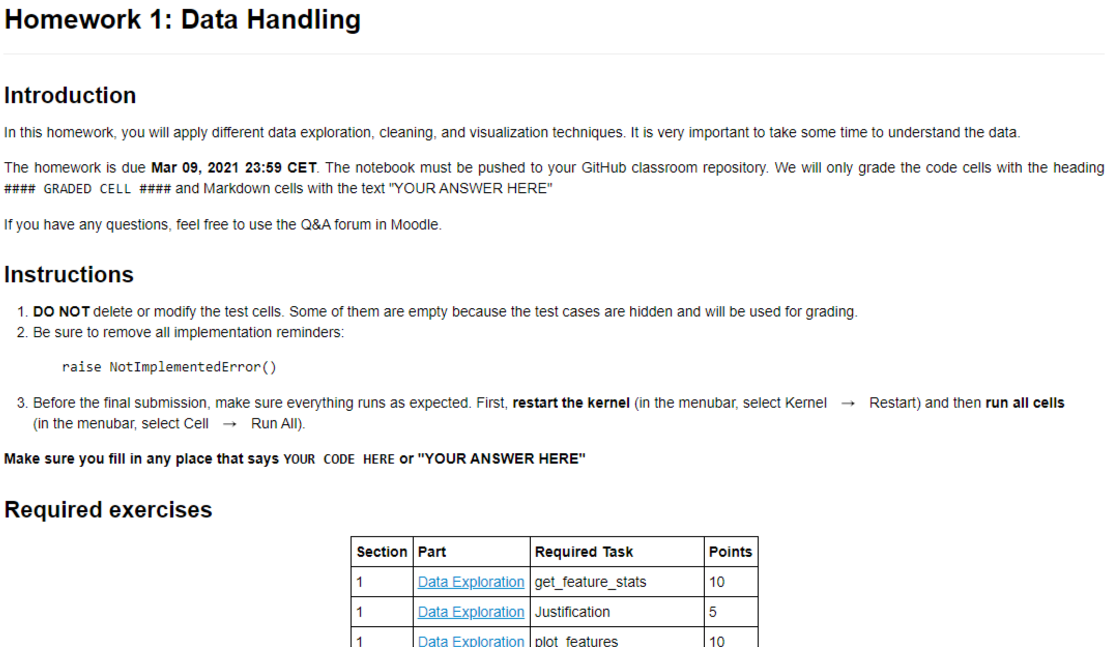
Feedback homework 1 (n=20)



8 Any additional comment?

Respondent	Response
	Some things were not very clear (e.g., what domains the data have, should we remove the missing/wrong values before displaying graphs etc)
	I am a totally beginner in pandas. So in my first homework, I have used many complex method to get what I want. I do not know if there exist a simple function to realize it.
	Couldn't finish it in time, thought 1 day would be enought but i didn't use pandas for a year and struggled with feature representation before cleaning.
	Weekly Labs and Homeworks combined gives us a huge amount of work and I already had to get myself late in other courses in order to finish them on time.
	I found the homework lengthy
	I spent 4hrs for homework and i didn't have enough time for exercises. I didn't give the best mark for the instructions because it was not really clear for me how can my code pass the tests
	I was disappointed with this homework. Too long and too unclear in my opinion.
	Continuing from 5): ...for instance, the exercise asking for descriptive statistics is not clear about the output df structure. If the tests allow for multiple structures, say so in the instructions
	The format of completing function cells is not ideal when working with datasets that you change and inspect all the time. I would prefer having more liberty with the notebook and we show the results.
	I'm worried about the workload. But it is useful too. Maybe make the tutorials shorter.

Extensive debriefing on homework 1



Homework 1: Data Handling

Introduction

In this homework, you will apply different data exploration, cleaning, and visualization techniques. It is very important to take some time to understand the data.

The homework is due **Mar 09, 2021 23:59 CET**. The notebook must be pushed to your GitHub classroom repository. We will only grade the code cells with the heading `#### GRADED CELL ####` and Markdown cells with the text "YOUR ANSWER HERE"

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

Instructions

1. **DO NOT** delete or modify the test cells. Some of them are empty because the test cases are hidden and will be used for grading.
2. Be sure to remove all implementation reminders:

   ```
   raise NotImplementedError()
   ```

3. Before the final submission, make sure everything runs as expected. First, **restart the kernel** (in the menubar, select Kernel → Restart) and then **run all cells** (in the menubar, select Cell → Run All).

Make sure you fill in any place that says `YOUR CODE HERE` or "YOUR ANSWER HERE"

Required exercises

Section	Part	Required Task	Points
1	Data Exploration	get_feature_stats	10
1	Data Exploration	Justification	5
1	Data Exploration	plot_features	10

The complete solution of this homework has been uploaded on the public GitHub repository at:

https://github.com/d-vet-ml4ed/mlbd/blob/main/Homework/Homework01/Homework01-DataHandling-Example-Solution-Complete.ipynb

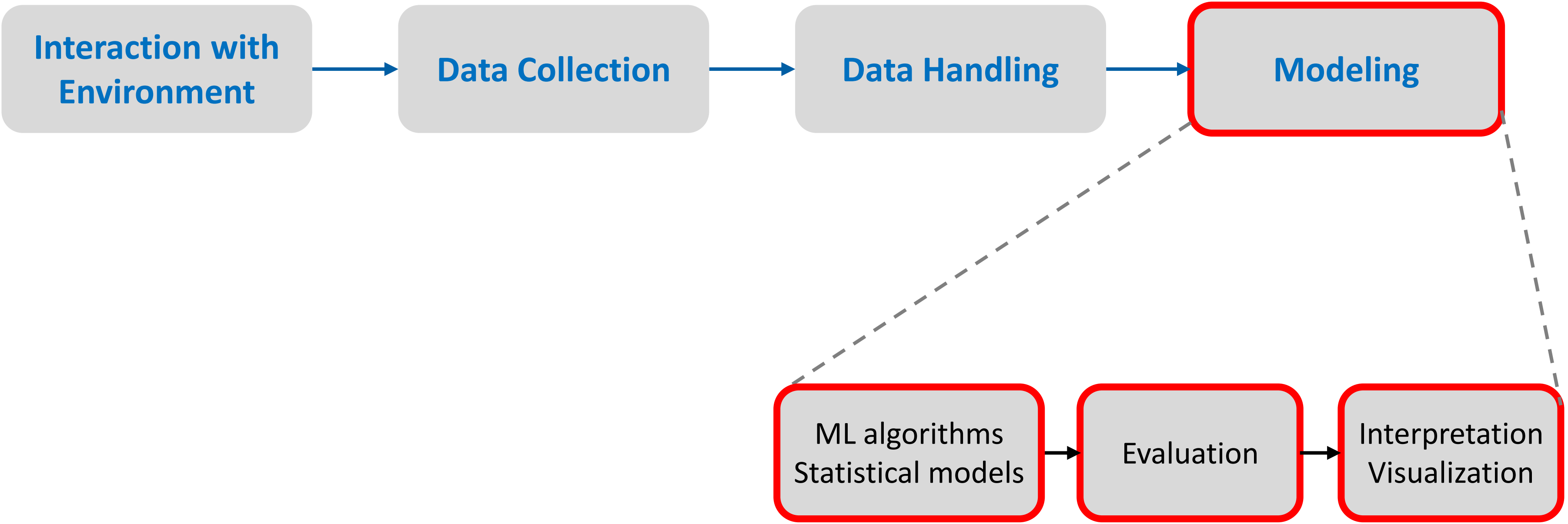
Questions?

Where we are

Week	Lecture	Lab Sessions	Project
1	Introduction	Tutorial	
2	Data Handling	Tutorial + **Homework**	
3	Regression & Classification	Tutorial + **Homework**	
4	Model Selection & Evaluation	Tutorial + **Homework**	Presentation of data sets and research questions
5	Latent Variable Models	Tutorial + **Homework**	M1: Preferences on team members and data sets
6	Unsupervised Learning	Tutorial + **Homework** + PO	
7	Spring Break	Spring Break	Spring Break

PO = project office hours

ML for Behavioral Data: Modeling



Interaction with Environment

Data Collection

Data Handling

Modeling

ML algorithms
Statistical models

Evaluation

Interpretation
Visualization

Tutorial 4.1 Agenda

- **15 mins** Fundamentals on model evaluation
  - Accuracy, Balanced Accuracy, Precision, Recall, F-Measure, Area Under ROC Curve, Confusion Matrix, Classification Report
  - Mean Absolute Error, (Root) Mean Squared Error, R^2
- **5 mins** Questions & answers time

Tutorial 4.1 Hands on

We will use Noto, but feel free to use your own environment:

- Go to https://noto.epfl.ch/
- Login with your GASPAR
- If you have **NOT** already cloned the repository:
  - Go to Git → Clone → https://github.com/d-vet-ml4ed/mlbd
- Go to Git → Pull
- Go through Tutorials/Tutorial04/Notebooks/Model_Evaluation.ipynb

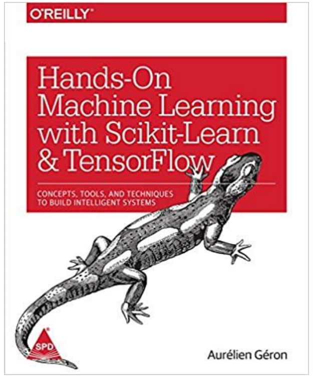
Questions?

Resources on model evaluation

Choose from below according to your **level**:

- https://scikit-learn.org/stable/modules/model_evaluation.html
  - Extensive description of the performance metrics.
  - More examples with several performance metrics.
- https://www.kdnuggets.com/2020/05/model-evaluation-metrics-machine-learning.html
  - Step-by-step explanation of performance metrics.
- Other interesting resources:
  - https://www.analyticsvidhya.com/blog/2020/10/how-to-choose-evaluation-metrics-for-classification-model/
  - https://www.kdnuggets.com/2018/06/right-metric-evaluating-machine-learning-models-2.html
  - https://ranvir.xyz/blog/how-to-evaluate-your-machine-learning-model-like-a-pro-metrics/

Further readings


O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
SPD
Aurélien Géron

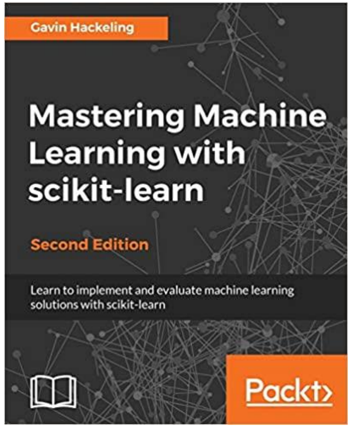
Section 2, Subsection "Select Performance Measure"

Section 3, Subsection "Performance Measures"

Géron, A. (2019). **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.** O'Reilly Media.

Section 4, Subsection "Performance Metrics"

Hackeling, G. (2017). **Mastering Machine Learning with scikit-learn.** Packt Publishing.


Gavin Hackeling
Mastering Machine Learning with scikit-learn
Second Edition
Learn to implement and evaluate machine learning solutions with scikit-learn
Packt>

Tutorial 4.2 Agenda

- **15 mins** Fundamentals on model selection
  - Holdout set strategies, cross validation strategies, bootstrapping
  - Validation curves, grid search and hyper-parameter tuning
- **5 mins** Questions & answers time

Tutorial 4.2 Hands on

We will use Noto, but feel free to use your own environment:

- Go to https://noto.epfl.ch/
- Login with your GASPAR
- If you have **NOT** already cloned the repository:
  - Go to Git → Clone → https://github.com/d-vet-ml4ed/mlbd
- Go to Git → Pull
- Go through Tutorials/Tutorial04/Notebooks/Model_Selection.ipynb

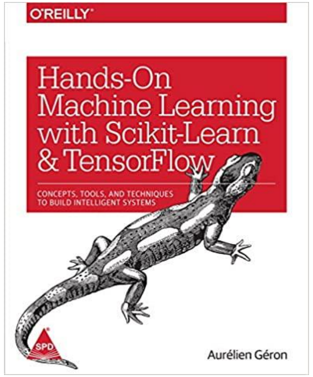
Questions?

Resources on model selection

Choose from below according to your **level**:

- https://scikit-learn.org/stable/model_selection.html
  - Extensive description of each evaluation method supported by scikit-learn.
  - Examples of toy datasets.
- https://www.kdnuggets.com/2019/10/choosing-machine-learning-model.html
  - Further details on elements that drive model selection.
- Other interesting resources:
  - https://towardsdatascience.com/a-complete-machine-learning-project-walk-through-in-python-part-two-300f1f8147e2
  - https://machinelearningmastery.com/nested-cross-validation-for-machine-learning-with-python/ (nested cross validation)

Further readings


O'REILLY
Hands-On Machine Learning with Scikit-Learn & TensorFlow
CONCEPTS, TOOLS, AND TECHNIQUES TO BUILD INTELLIGENT SYSTEMS
SPD
Aurélien Géron

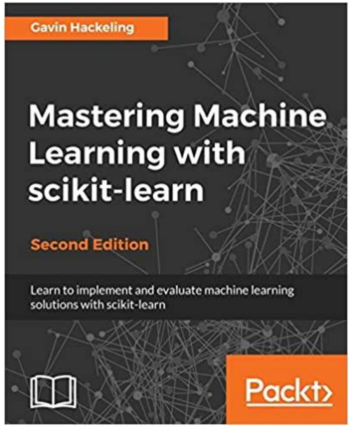
Section 2, Subsection "Select and Train a Model"

Section 3, Subsection "Fine-tune Your Model"

Géron, A. (2019). **Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems.** O'Reilly Media.

Section 1, Subsection "Training data and test data"

Hackeling, G. (2017). **Mastering Machine Learning with scikit-learn.** Packt Publishing.


Gavin Hackeling
Mastering Machine Learning with scikit-learn
Second Edition
Learn to implement and evaluate machine learning solutions with scikit-learn
Packt

Class Project

- **Rules**:
  - You must create a team of 3 people
  - You will work on one of the four tracks we provide (see **project descriptions**)
  - You will work on one of the datasets we provide (see **project descriptions**)
  - We will provide **example** research tasks (see **project descriptions**)
- **Workflow**:
  - You will work on the project, from research questions to models and results
  - We will do project office hours (during lab **sessions**)
  - We will give feedback during the semester (see **milestones**)
  - You will do a short presentation in the last week of the semester
  - You must submit code + report by June 11, 2021 23:59 CET (see **project descriptions**)

Track #1 – E-Tutoring

EdNet Datasets or PSLC DataShop Datasets, up to 6GB

These repositories include datasets with student's actions, e.g., material they consumed, response, how much time they spent on a given question or reading expert's commentary.

Example tasks:

- Predicting student response correctness (correct/incorrect) to newly-encountered questions.
- Predicting the probability that a student responds correctly to newly encountered multiple-choice questions.
- Investigating the study session dropout prediction problem in mobile learning environments.
- Devising an approach for knowledge tracing along the interactions of students in the platform.
- Creating models that can provide learning path recommendations to each student.
- Identifying and describing groups of students characterized by similar learning patterns.
- Investigating the behavioral features that lead to successful learning outcomes by students.

Track #2 – E-Commerce

Amazon Datasets (only small data), < 1GB or up to 6GB

These Amazon review datasets include reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs), in addition to product metadata.

Example tasks:

- Predicting the list of products or the next product to recommend to each user.
- Predicting how much a review is helpful for people (e.g., helpful/unhelpful or a helpfulness score).
- Predicting the satisfaction of a reviewer (i.e., the rating) based on the review text.
- Predicting the list of products to suggest to each user based on multiple sources (e.g., ratings, reviews).
- Segmenting customers according to their preferences and online behavior.
- Investigating the extent to which the popularity of a product or the market in general changes over time.
- Modelling the different aspects of a product review and clustering reviews accordingly.

Track #3 – Music

LastFM Datasets, up to ~6 GB

These datasets track user activity in music-centric applications and include, for instance, <user, artist-mbid, artist-name, total-plays> tuples representing the interactions of users.

Example tasks:

- Predicting the next song to suggest to a user, when they play with a music platform.
- Identifying groups of individuals who share similar song listening patterns.
- Predicting a playlist that satisfies group members (e.g., to decide the music to play in a party).
- Predicting to what extent a song will become popular over time and what drives popularity.
- Identifying polarities or filter bubbles based on the song recommendations provided to users.
- Studying the consumption of music across years and genres.

Track #4 – E-Learning

Xuetang Datasets, up to ~2GB

XuetangX is a massive open online course (MOOC) platform that offers online courses in multiple disciplines. These datasets include student profile information, course metadata and descriptive attributes, student learning interactions and feedback (e.g., ratings).

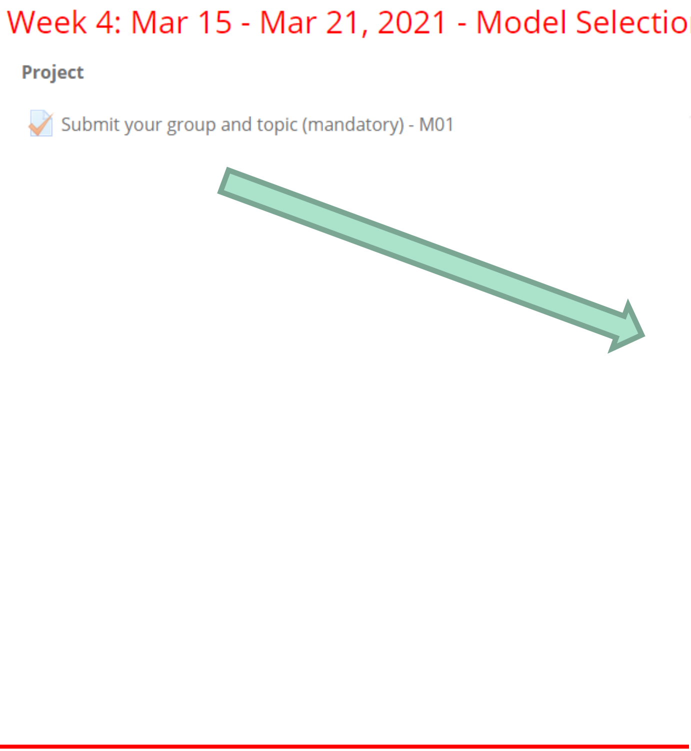
Example tasks:

- Predicting the next course to recommend to each student, based on their needs or interests.
- Predicting whether a student will drop out of a course, based on their interactions in the platform.
- Analyzing factors that influence students’ engagement and how they relate to learning effectiveness.
- Identifying the relationships across courses, in order to shape meaningful learning course paths.
- Predicting knowledge concept recommendations to students based on their past behavior.

Milestones

- **M01** on Preferences on Tracks and Group Members
  - due March 23, 2021 23:59 CET - **MANDATORY**
- **M02** on Research Questions and Exploratory Analysis
  - due April 13, 2021 23:59 CET - optional
- **M03** on Implemented Approach and Preliminary Results
  - due May 04, 2021 23:59 CET - optional
- **M04** on Mature Approach and Results with Discussion
  - due May 18, 2021 23:59 CET – optional
- **Project Presentation** for Course Evaluation
  - to be given on May 31, 2021 - **MANDATORY**
- **Final Project Deliverable** for Course Evaluation
  - due June 11, 2021 23:59 CET - **MANDATORY**

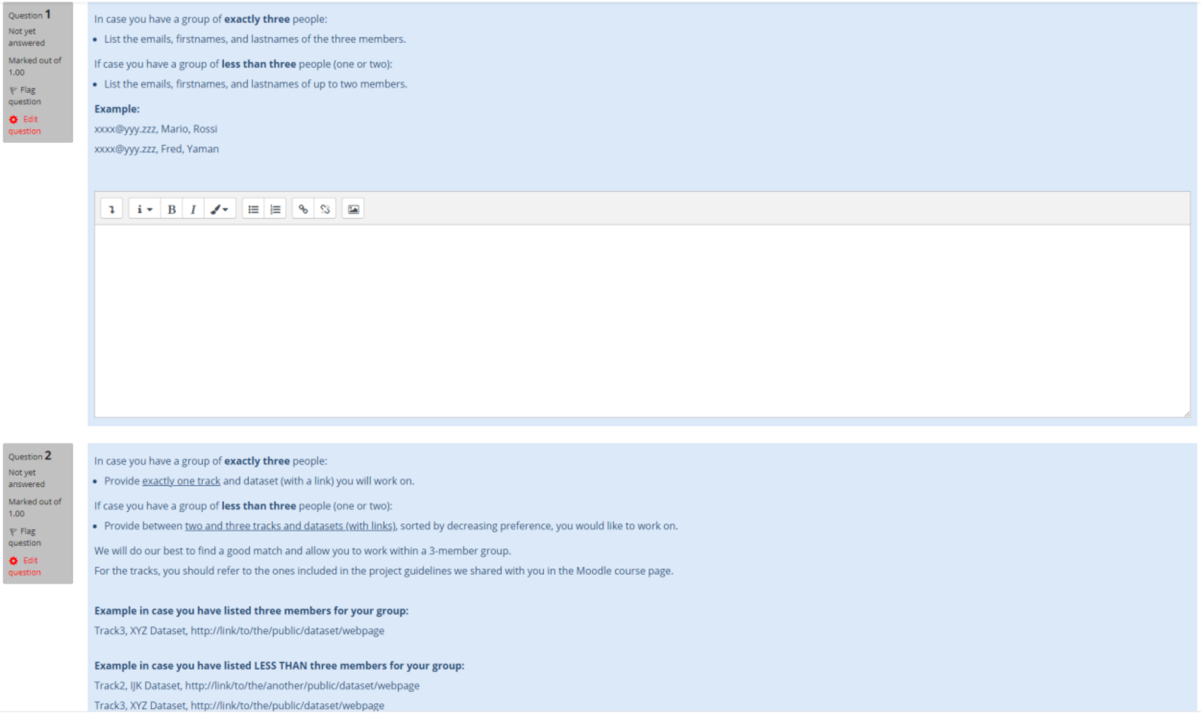
First project milestone (mandatory)



Week 4: Mar 15 - Mar 21, 2021 - Model Selection & Evaluation

Project

Submit your group and topic (mandatory) - M01



In case you have a group of **exactly three** people:

- List the emails, firstnames, and lastnames of the three members.

If case you have a group of **less than three** people (one or two):

- List the emails, firstnames, and lastnames of up to two members.

Example:

xxxx@yyy.zzz, Mario, Rossi

xxxx@yyy.zzz, Fred, Yaman

In case you have a group of **exactly three** people:

- Provide exactly one track and dataset (with a link) you will work on.

If case you have a group of **less than three** people (one or two):

- Provide between two and three tracks and datasets (with links), sorted by decreasing preference, you would like to work on.

We will do our best to find a good match and allow you to work within a 3-member group.

For the tracks, you should refer to the ones included in the project guidelines we shared with you in the Moodle course page.

Example in case you have listed three members for your group:

Track3, XYZ Dataset, http://link/to/the/public/dataset/webpage

Example in case you have listed LESS THAN three members for your group:

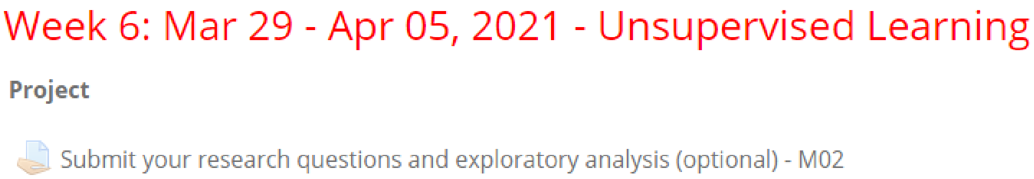
Track2, IJK Dataset, http://link/to/the/another/public/dataset/webpage

Track3, XYZ Dataset, http://link/to/the/public/dataset/webpage

Other optional project milestones

- You could prepare up to 2 pages with descriptions and artifacts you worked on.
- Submit the corresponding file into Moodle by the milestone deadline.

For instance, for M02 “Research Questions and Exploratory Analysis”, you could prepare 2 pages that include the research questions and a series of artifacts for your exploratory analysis. Decisions must be motivated and discussed, as in HWs.



Week 6: Mar 29 - Apr 05, 2021 - Unsupervised Learning

Project

Submit your research questions and exploratory analysis (optional) - M02

(available soon)

Final deliverable (mandatory)

- By June 11, 2021 23:59 CET, you will need to deliver a written report of your project that will be of **10 pages at maximum**.
  - ACM's Template Word or Latex
  - Suggested outline with Introduction, Exploratory Analysis, Approach, Experimental Evaluation, Discussion, Conclusions and Future Works.

- What you need to submit:
  - Written report (pdf)
  - Final presentation (pptx or any other editable format)
  - Codebase (zip file)

Grading

- The class project accounts for 40% of the course grade.

- The class project grade will be evaluated as follows:
  - Final presentation 30%
  - Report and Codebase 70%

Class project description

For an extensive description of the project details, please go over this document:

https://tinyurl.com/2021-CS-421-Project

This link will be present also on Moodle, under the “Week 4” section!

Class project ethics

For considerations on ethics on machine learning for behavioral data, please go over this document:

https://tinyurl.com/Ethics-Considerations-MLBD

This link will be present also on Moodle, under the “Week 4” section!

Introduction to homework 3

- In **telecommunication companies**, different reasons trigger customers to **terminate** their contracts, such as better price offers, more interesting packages, …
- Churn analytics provides valuable capabilities to predict **customer churn** and define the underlying reasons that drive it.
- We will imagine that a company asks you to predict whether its customers are about to churn by means of a **machine-learning approach**. Specifically, we will ask you to:
  - Load, explore, and pre-process data appropriately, based on the task and application (what is important?).
  - Implement and evaluate a machine-learning approach appropriately, based on the task and data set (e.g., size).

Important upcoming dates

- Mar 22, 2021 15:15 17:00
  - Lecture #5
- By Mar 23, 2021 23:59 CET
  - Submission Deadline for Homework #3
- By Mar 23, 2021 23:59 CET
  - Submission Deadline for Project Preference on Members and Tracks
- Mar 24, 2021 08:15 10:00
  - Lab Session #5

Your feedback is essential

- It is a new course, please give feedback on how to improve it
- Short anonymous feedback forms on Moodle

https://moodle.epfl.ch/mod/questionnaire/view.php?id=1133027

Quick Anonymous Feedback on Lab Session 4

Questions?