



*Faculty of Mechanical Engineering
and Robotics*

*Department of Robotics and
Mechatronics*



Python for Machine Learning and Data Science *Course for Mechatronic Engineering*

Instruction 2:

Project datasets introduction

You will learn: How to start work with the unknown datasets. How to formulate tasks and research questions.

Additional materials:

- Course lectures 1 & 2 [*obligatory*]
<http://galaxy.agh.edu.pl/~zdw/Materials/Python/LectureNotes/>
- Report template
<http://galaxy.agh.edu.pl/~zdw/Materials/Python/>

Learning outcomes supported by this instruction:

IMA1A_U01, IMA1A_U05, IMA1A_U06, IMA1A_U07, IMA1A_K08

Course supervisor:

Ziemowit Dworakowski, zdw@agh.edu.pl

Instruction author:

Adam Machynia, Ziemowit Dworakowski, machynia@agh.edu.pl

Introduction

This laboratory will use knowledge from the previous laboratory to develop the initial section of the project tasks. Note that this laboratory is designed to be completed in teamwork – you should divide responsibilities for particular tasks among different team members and coordinate your work. Some tasks can be done in parallel, while others serve as prerequisites for the latter part of the instruction.

You do not receive a mark for each class devoted to the project; however, there are tasks that are mandatory to pass a particular lab. You must always complete the "red" tasks during the lab. "Orange" tasks are obligatory for your project, so you should finish them during the lab or complete and present them later. Finally, "green" tasks are optional and will increase your project grade. Therefore, if you only complete the red and orange tasks, you will aim for a 4.0 mark for your project.

Warm-up (30 – 45 minutes)

At the beginning, we will use the following dataset from Kaggle:

<https://www.kaggle.com/datasets/aakashjoshi123/exercise-and-fitness-metrics-dataset/>

Your task will be to prepare a rough analysis of this dataset and draw some initial conclusions.

Task 2.1: Download and load the *Exercise and Fitness Metrics Dataset* into the Colab. Take a look at the dataset and its features. Prepare some basic plots and check the distribution of the features. Discuss your observations with your team (all team members should contribute to this task). Be ready to present your conclusions and discuss them with the teacher.

Context and related work

Your project task consists of a specific dataset, typically acquired to address a particular scientific problem. Read about the contents of the dataset in its description on the UCI MLR website or Kaggle. Note that datasets from UCI MLR are generally also available on Kaggle. Consider the task ahead. Are there any biases you can identify, or general-knowledge-based assumptions about the dataset that you already have? Possible examples for different datasets may include, for instance, the assumption that poor diet or drug use increases the likelihood of health deterioration; the assumption that good weather leads to more outdoor workouts; or that older cars tend to have higher mileage.

Task 2.2: Biases and assumptions
Prepare a list of possible biases or assumptions you have as you begin this task. Discuss this issue within your team (all team members should contribute to this task). Store this list for further use and interpretation of results in the latter part of the project.

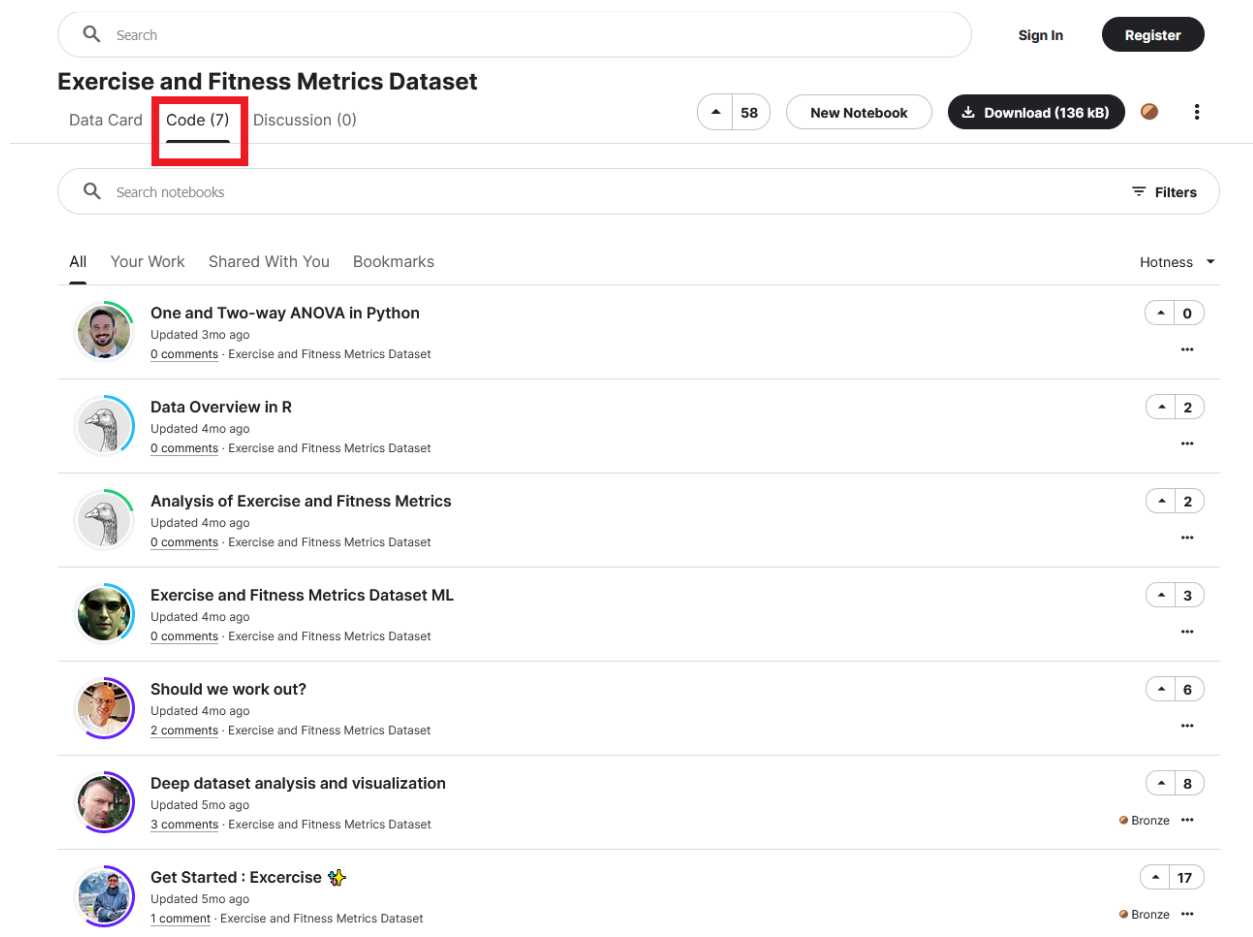
To provide context for your work, you should know the source of the data and what work has been done previously on this dataset. For instance, look at the dataset provided for the warm-up. In the Code section, you will see examples of tasks already solved by others (see Figure 1). Also, check the Metadata section of the Data Card, where you can find information about related papers or detailed research descriptions that will help you put the dataset into context.

Task 2.3: Related work

Look for information about your project dataset. Enumerate all the work done previously on this dataset by others, describe its contents (what was done, what libraries or programming languages were used, the goals of the analyses, and the general conclusions obtained by the authors).

Task 2.4: Prepare a report on the results of Task 2.3. This report will serve as the second chapter of your project report. It should be prepared in a text editor or in LaTeX (a Python notebook is not an acceptable form of a report).

You are encouraged to use generative AI for this task. However, make sure to store the prompts used for this purpose. You should prepare an appendix to the report where you describe how you obtained this chapter. Feel free to discuss the use of prompts with the teacher.



The screenshot shows the Kaggle interface for the 'Exercise and Fitness Metrics Dataset'. At the top, there is a search bar and buttons for 'Sign In' and 'Register'. Below the dataset title, there are tabs for 'Data Card', 'Code (7)', and 'Discussion (0)'. The 'Code' tab is selected and highlighted with a red box. To the right of the tabs, there are buttons for 'New Notebook' and 'Download (138 kB)'. Below the tabs, there is a search bar for notebooks and a 'Filters' button. The main content area shows a list of notebooks related to the dataset, each with a profile picture, title, update time, and comment count. The notebooks listed are:

- One and Two-way ANOVA in Python (Updated 3mo ago, 0 comments)
- Data Overview in R (Updated 4mo ago, 0 comments)
- Analysis of Exercise and Fitness Metrics (Updated 4mo ago, 0 comments)
- Exercise and Fitness Metrics Dataset ML (Updated 4mo ago, 0 comments)
- Should we work out? (Updated 4mo ago, 2 comments)
- Deep dataset analysis and visualization (Updated 5mo ago, 3 comments)
- Get Started : Exercise (Updated 5mo ago, 1 comment)

Fig 1 - Dataset code section.

Dataset loading and division

You will be using the procedures described in the first instruction to load and utilize your project datasets. Remember that in all data science projects, it is essential to reserve a portion of the dataset for final evaluation to allow for a blind test of the final set of hypotheses. For this reason, before you start analyzing the data, you should randomly select a subset that will not be used for data visualization, hypothesis development, testing, and AI model implementation and configuration.

Task 2.5: Dataset loading

Download your project dataset, store it in your Google Drive, and then connect it to your Colab account. Show the description of the dataset to ensure its correct connection to your code.

Task 2.6: Dataset division

Divide the dataset into subsets. Randomly reserve 30% of the dataset for testing purposes at the end of the project. All further tasks in this instruction should be performed only on the remaining 70% of the data samples.

Dataset visualization and research hypotheses

During the first laboratory, you learned about various tools that can be used to visualize your datasets. You should take into account the following questions:

- 1) How is the data divided among different categories (using binary and polytomous categorical variables)? Are there any categories that have relatively few samples?
- 2) How are numeric features distributed? Is this distribution similar among different categories?
- 3) Are there clear relationships between different features that can be inferred straight from data?
- 4) What are the possible sources of the observed correlations?
- 5) Are there any interesting research questions and observations that can be derived from the data?
- 6) Are there features that do not seem important in the context of potential problems?

Task 2.7: Dataset visualization

Examine your dataset, perform exploratory data analysis, and show how features are distributed and their relationships to each other. Prepare a Python notebook that includes the necessary figures.

Task 2.8: Project task formulation

Based on your observations, formulate research questions that will serve as potential goals for your project. Try to prepare several potential tasks and then discuss them with your teacher. Note that these tasks should not repeat those already completed in the related work (Task 2 should be done prior to this one).

To complete this task, you will need at least one classification or regression problem. After receiving approval from the teacher, these research questions will form the final goal of your project.

Task 2.9: Hypothesis formulation

Based on your observations and the previous task, try to prepare several potential hypotheses. Formulate them clearly and consider what should be done to prove or reject them. Discuss your ideas with the teacher and decide together if these will become part of your project.

Task 2.10: Exploratory data analysis report

Using the materials prepared in Task 2.7, prepare a document that will form chapter 3 of your project report. This document should be prepared in a text editor or in LaTeX (a Python notebook is not an acceptable form of a report). Provide an explanation of the steps taken and an interpretation of the obtained results.

Business presentation

The last part of this instruction consists of hints and tools regarding the business presentation that will be prepared for project classes based on the results obtained in this laboratory. The goal of this step is to get familiar with typical way of presenting your projects and accomplishments in a short and concise form, focusing on a story and powerful idea instead of complex details and background. The presentation will last up to 3 minutes with an additional follow-up discussion. Such short time reflects live situations when you might be given a chance to interest someone with your concepts during a short meeting, talk during lunch break or a discussion panel.

Things worth considering:

1. 3 minutes is much shorter than you think. Each second is precious and should be used consciously. You won't have time to show a detailed background or explain details of method's application. Ensure that you actually trained your speech – you will be asked to stop after the timer runs out.
2. Think of a powerful start to your presentation. You aim to catch attention of the audience. Make it interesting. Consider starting from a question, interesting fact or a seemingly contradictory sentence.
3. Don't memorize the whole script you want to say. You will sound artificial (it takes a great actor to actually say scripted statements naturally). Instead memorize keywords that you want to use in each sentence and a structure that you want to follow.
4. Don't focus on slides too much. Reading from slides sometimes makes a somewhat decent lecture, but very poor business presentation. You want to have up to 5 slides with few sentences on them (the more text on the slides, the worse presentation becomes)
5. Consider presenting entirely without slides. In 3 minutes you can often make a better impression without using power point at all. Lack of slides allows usually for better contact with audience and less distractions.
6. Don't overcomplicate and don't try to assign roles for many team members (like: different people explain different things) unless you have a really good idea on how to do it fluently. This usually breaks the flow of a talk and renders the audience confused. "One person talking" is a recommended approach here.
7. Build a story. Your goal is to provide narration that could get you funding for your work or research. You want to convince audience that your goals are really interesting, not just reasonable to follow. Emotionally invested audience is already half of the success in further business negotiations.

8. Say why this project is unique. Comment (very briefly) on what others have done and explain why your approach is better (or at least: addresses things in a different way).
9. Say why this project is doable. Explain what relations in data led you to believe that you will have a working solution.
10. End on a powerful note. Think big, You don't want to pursue 'a better classification accuracy'. You want to actually make a significant change with your research. Explain how the project results will alter what we know and do today.
11. Do a rehearsal with other team members at least once. Ask other team members to assume critical position on your talk, ask them to pinpoint any issues, changes in talk dynamics, question motivation, goal and feasibility of the project (in relation to your talk).

Task 2.11: Business presentation

Prepare a 3-minute talk that will show your project's goal and explain motivation for it. Pretend that after initial data analysis you look for funding of an external investor to complete the project. You can (but you don't have to) use slides. This presentation will be given during next project classes.