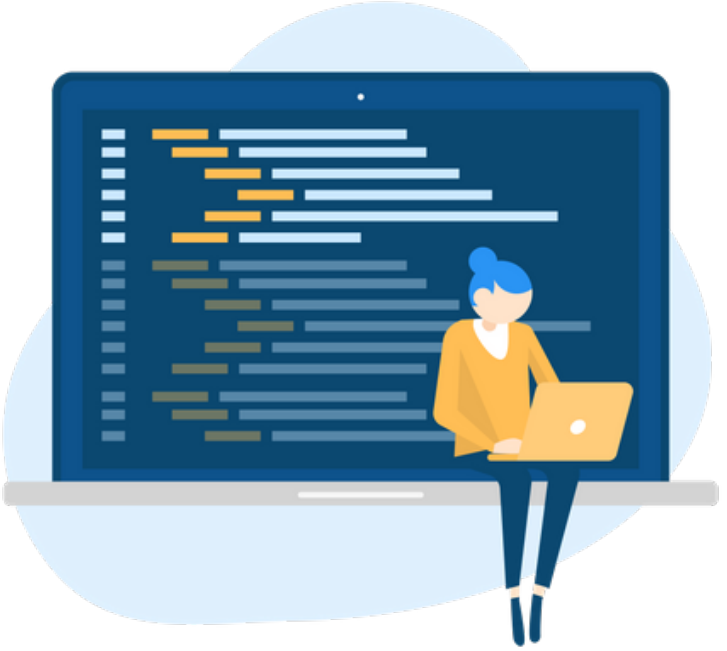
POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Final Project

Programming for Data Science



ETSI Informáticos

Universidad Politécnica de Madrid

General aspects

1. Groups of 4 students
2. One final project with two parts (but just one submission):
   - Submission deadline: by 10/04/2024
   - Moodle task for project submission
3. Video recording to present work developed in the final project

- Just one submission by group of students
- **IMPORTANT**: During classes you can ask for doubts. Also, tutoring time for doubts can be appointed. But, **no doubts will be solved by email or Moodle Forums**.

POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Data collection

First part

ETSI Informáticos

Universidad Politécnica de Madrid

Data Collection

1. **Free topic**

   Each group of students can choose the web site to access and the data to extract from the selected web site. Main requirement is that information must be retrieved from the web site using Web Scraping (static or dynamic) and store the extracted data in tabular format (csv).

   Decide about a topic of your interest and implement a script that interacts with the selected web site, according to the chosen topic of interest.

   Examples: financial data (investing.com), scientific publications and authors (Google scholar, pubmed, dblp…), Data about movies (imdb), …

2. **Requirements**
3. **Expected result**

Data Collection

1. **Free topic**
2. **Requirements**

Following requirements to take into account:

- It should be extracted at least 3 resources from the same web site
  - What is a resource? an entity with associated information. Example: in the finance use case with 'investing.com', a resource could be the stock of a company, and the information to retrieve from the resource (in that case) would be the daily price of the company stock during the last two years, for doing later an analysis of time evolution of stock price (return, risk, volatility, …)
- Handle correctly implicit/explicit waiting for loading/accessing web content
- Automatic interaction with the web site (i.e. load URL, click on links/buttons, extract and store data, …)

IMPORTANT: Avoid websites hard to process. Be clever to choose a website that your team can process and extract data, no matter if it is simple and easy to scrap. It is more valuable good analysis questions well answered, even if data used was easily obtained.

3. **Expected result**

Data Collection

1. **Free topic**
2. **Requirements**
3. **Expected result**

The developed script should extract from the chosen web site, the selected information. Extracted information should be stored in tabular format (.csv), one .csv file per extracted resource.

Example: in the finance example (investing.com) the temporal data relative to different types of resources or financial assets (e.g. company stock, corporate bonds, bitcoin, …) should be stored in different .csv files. One .csv file per asset or resource.

So, following outcomes expected for first part:

- Python script to retrieve data from web site and store it in .csv files
- The different .csv files obtained *

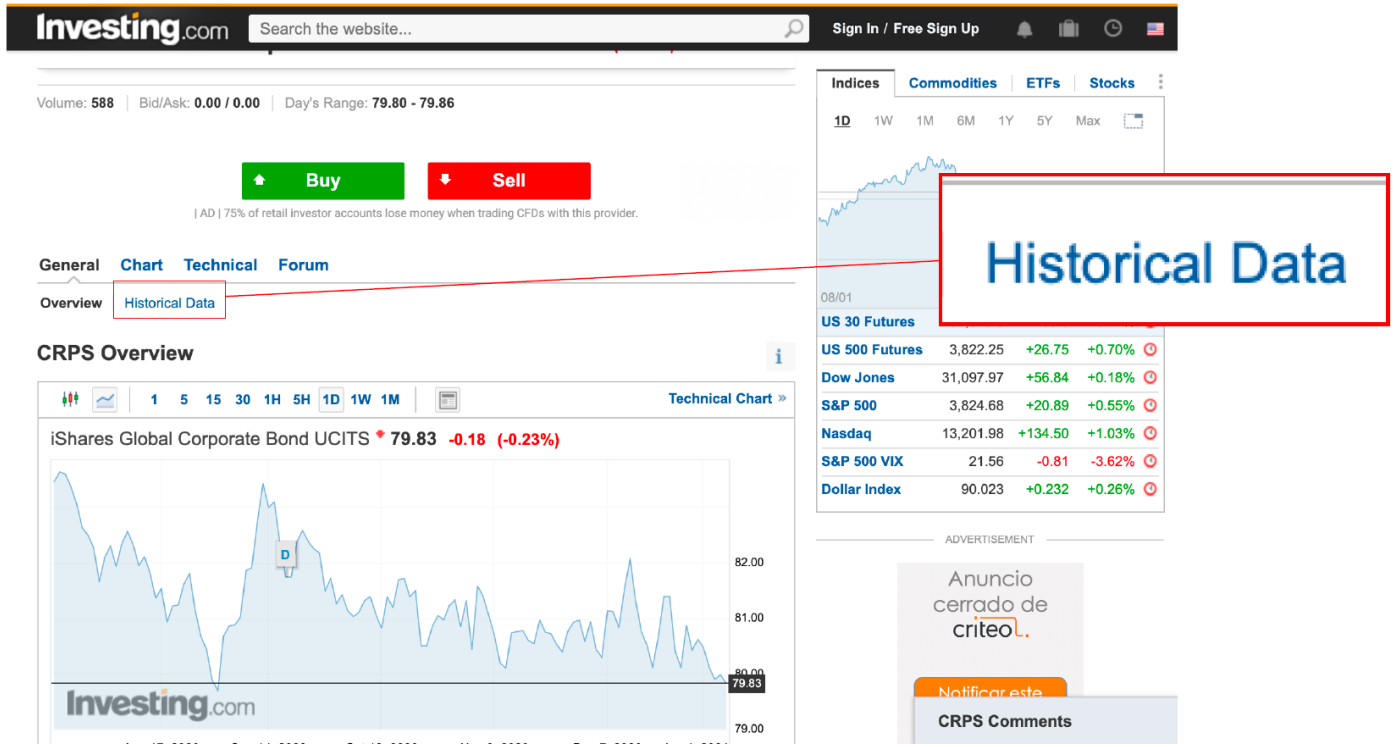
* These files will be the input for second part of the final project

Example: Data Collection

- Data collection about financial assets from investing.com
- The purpose of the example is to retrieve historial data about different financial assets:
  - Amundi Index Msci World Ae-c, https://www.investing.com/funds/amundi-msci-wrld-ae-c
  - iShares Global Corporate Bond (CRPS), https://www.investing.com/etfs/ishares-global-corporate-bond-$
  - Xtrackers II Global Government Bond, https://www.investing.com/etfs/db-x-trackers-ii-global-sovereign-5
  - SPDR® Gold Shares (GLD), https://www.investing.com/etfs/spdr-gold-trust
  - US Dollar Index, https://www.investing.com/indices/usdollar
- By dynamic interaction with the web site, historical data can be retrieved by inserting dates in some form fields and submitting the query form (see later).

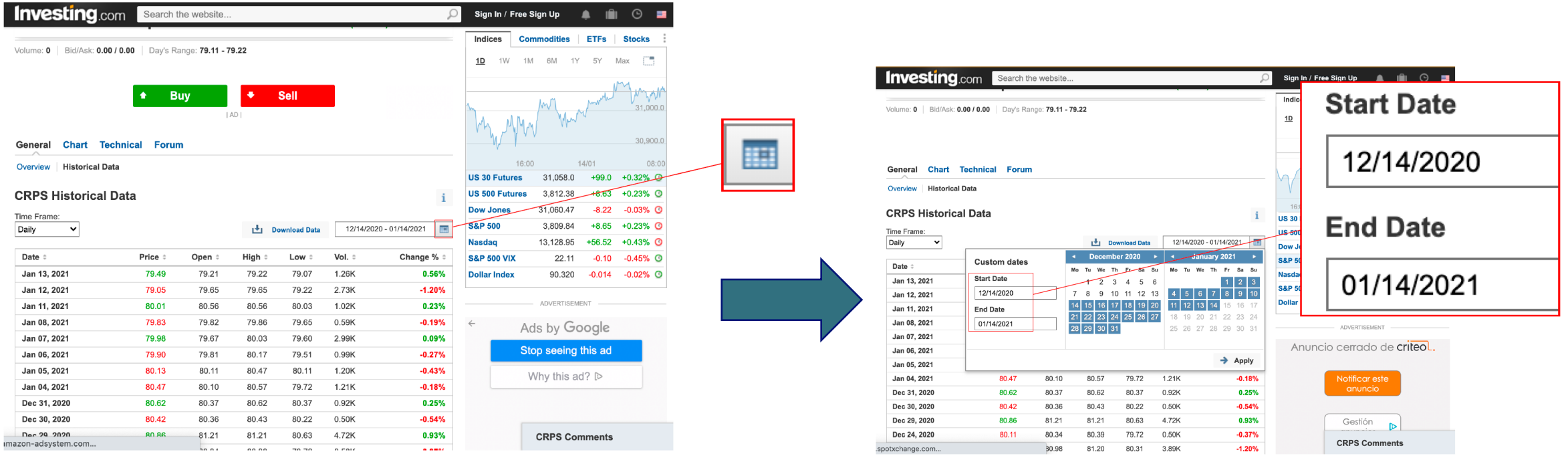
Example: Data Collection

- Main page of a given financial asset
  - By clicking on ‘Historical Data’ link we can obtain page with asset historical price data (see next slide)



Example: Data Collection

- Asset historical data
  - Page with table of asset prices over time and calendar button to select date ranges (see left)
  - Floating form opened to insert in text fields the date ranges for data retrieval (see right)



Example: Data Collection

- Asset historical data collection
  - Process the table in the asset page to extract data about asset price evolution over time

Date	Price	Open	High	Low	Vol.	Change %
Jan 13, 2021	79.49	79.21	79.22	79.07	1.26K	0.56%
Jan 12, 2021	79.05	79.65	79.65	79.22	2.73K	-1.20%
Jan 11, 2021	80.01	80.56	80.56	80.03	1.02K	0.23%
Jan 08, 2021	79.83	79.82	79.86	79.65	0.59K	-0.19%
Jan 07, 2021	79.98	79.67	80.03	79.60	2.99K	0.09%
Jan 06, 2021	79.90	79.81	80.17	79.51	0.99K	-0.27%
Jan 05, 2021	80.13	80.11	80.47	80.11	1.20K	-0.43%
Jan 04, 2021	80.47	80.10	80.57	79.72	1.21K	-0.18%

POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Data Manipulation and Visualization

Second part

ETSI Informáticos

Universidad Politécnica de Madrid

Data Manipulation and Visualization

1. **Topic of the second part**

Using the data obtained in the first part using Web Scraping, this part is about processing and visualizing the data in order to perform some analysis.

The analysis should be performed in order to answer a set of questions related with the data. The questions must be defined by each work group and should be answered providing support in the form of data (lists, tables, …) or plots.

It will assessed the creativity and the interest of questions formulated and answered.

2. **Data treatment**
3. **Data visualization**

Data Manipulation and Visualization

1. **Topic of the second part**
2. **Data treatment**

It must be performed at least one data transformation of the methods and techniques covered in the course (i.e. manipulation, wrangling and/or transformation).

Examples:

- Group research articles from pubmed by topic or by keywords or by authors belonging to the same research team
- Group movies by theme/topic or by actors or by producers … etc

3. **Data visualization**

Data Manipulation and Visualization

1. **Topic of the second part**
2. **Data treatment**
3. **Data visualization**

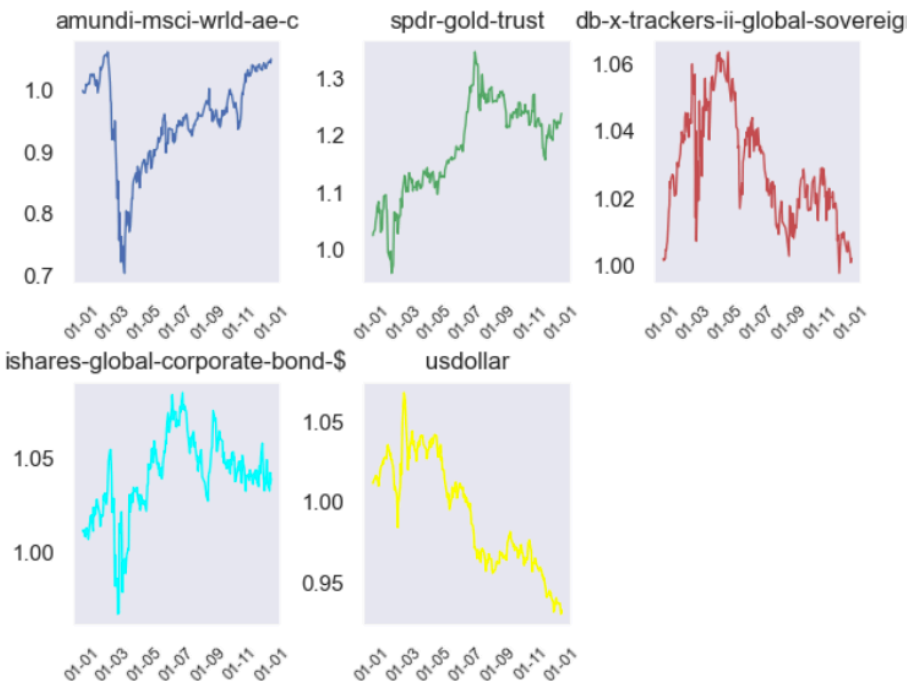
The part of data visualization should be guided by the questions to answer. Thus, each work group should define a set (i.e. 2-3) of questions to be answered providing support in the form of plots (covered in the course) and/or any other type of information extracted from collected data.

Examples: In the case of movies, answer questions like: which are the themes/topics of the top-k most awarded movies? Or top-k list of movies with good valuation from audience but with reduced or no prizes …etc

Example: Data Collection

- Data collection about financial assets from investing.com
- Historical traded prices in markets for 5 different assets:
  - Amundi Index Msci World Ae-c, https://www.investing.com/funds/amundi-msci-wrld-ae-c
  - iShares Global Corporate Bond (CRPS), https://www.investing.com/etfs/ishares-global-corporate-bond-$
  - Xtrackers II Global Government Bond, https://www.investing.com/etfs/db-x-trackers-ii-global-sovereign-5
  - SPDR® Gold Shares (GLD), https://www.investing.com/etfs/spdr-gold-trust
  - US Dollar Index, https://www.investing.com/indices/usdollar

- By web scrapping, interacting with the website, is obtained the temporal evolution of asset prices.



Example: Data manipulation and visualization

- From the collected data, up to 126 different investment portfolios are generated. Each portfolio defines asset allocation with different weights for each of the available assets:
  - ST: Stocks
  - CB: Corporate Bonds
  - PB: Public Bonds
  - GO: GOld
  - CA: CAsh

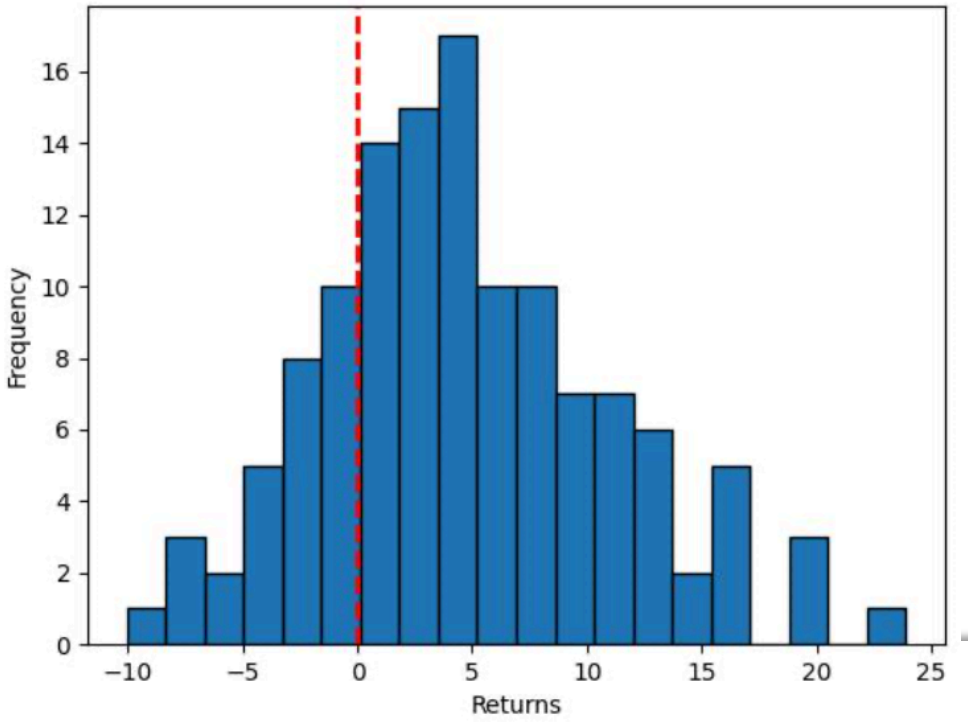
Asset Alloc.	ST	CB	PB	GO	CA
1	100%	0%	0%	0%	0%
2	80%	20%	0%	0%	0%
3	80%	0%	20%	0%	0%
4	80%	0%	0%	20%	0%
5	80%	0%	0%	0%	20%
6	60%	20%	20%	0%	0%
7	60%	20%	0%	20%	0%
		...			
i	40%	60%	0%	0%	0%
i+1	40%	0%	60%	0%	0%
i+2	40%	0%	0%	60%	0%
i+3	40%	0%	0%	0%	60%
		...			
i+n	0%	100%	0%	0%	0%
		...			

Example: Data manipulation and visualization

- Once generated the different investment portfolios, the 12 months return of each portfolio is computed taking into account the buy Price and sell Price for each asset weight.
- Correct treatment of missing values should be decided. In this case, missing values means market are closed, so if missing prive value -> keep the price from the previous available day.
- Example question:

"Taking into account the 12 month return of ALL (i.e. 126) investment portfolios. Is it more probable to obtain positive or negative return?"

Yes, looking at the plot we can say it is more probable to obtain positive returns because …



Presentation

Each group must prepare a video recording of around 10 minutes explaining the two parts of the final Project implemented.

For the video recording a set of slides can be prepared to help the presentation of the group work. In these slides it should be BRIEFLY explained everything:

- First part (Web Scraping)
  - Web site(s) selected
  - Resources chosen
  - Explain how scrapping is performed (at high level)
  - Files generated
- Second part (Data manipulation and visualization)
  - Questions defined for the input data
  - Steps to answer for each question:
    - Data tasks (transformation, manipulation, …)
    - Data visualization
    - Answer to question with support material (plots, …)

REMEMBER!! Presentation (using slides) should last 10 minutes at MOST! It should not delve too much into low level details of implementation. Everything should be explained at high level (e.g. text and diagrams, …). If necessary, showing some code in the slides is allowed but do not abuse …

Submission

A .zip file should be submitted containing:

- Name your .zip file as: pdp_group-<group_id>.zip (e.g. pdp_group-6.zip)
- Python files with your script(s), do not forget to document properly the code.
- README.txt file containing:
  - List of students that form the group
  - Description of files (and folders if any) with a brief explanation about how to execute the code. It should be also specified any package dependencies. In case of splitting the project in several modules, briefly indicate a description of each module.
  - A link to watch the video recording presenting the final project (see previous slide)
- The slides (in pdf) used in the video recording for presenting implementation and results of the final project

Submission

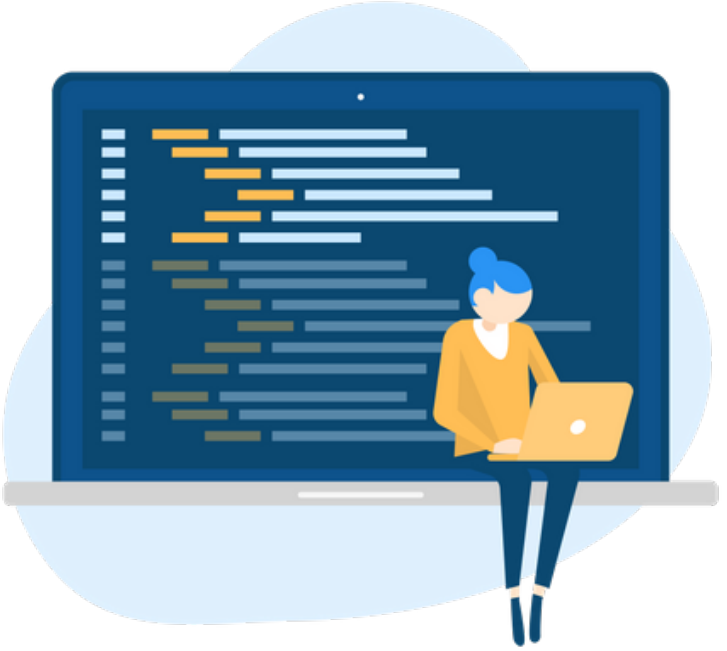
Submission deadline: 10/04/2024
One submission per group of students
Submission through a Moodle task

POLITÉCNICA
"Ingeniamos el futuro"

CAMPUS
DE EXCELENCIA
INTERNACIONAL

Final Project

Programming for Data Science



ETSI Informáticos

Universidad Politécnica de Madrid