Programming for Data Science 2025

Guidelines

As a Data Scientist, your role is to bring valuable insights into problems by relying on data. To do so, you must acquire, filter, clean, and merge data from multiple sources. You must also classify information, cluster or segment data, and identify non-intuitive relationships between features. Additionally, you will prepare meaningful visualizations to communicate your findings and construct a narrative to report your results.

In this project, your main goal is to explore a problem of interest using Python as your primary tool. As this is your first semester in the program, we want you to focus on data acquisition, data cleaning, and the description/exploration of a phenomenon. Therefore, modeling aspects such as model complexity or performance will not be considered for grading. Instead, we want you to emphasize the effort on data cleaning, description/exploration, and visualization.

The project is an open-topic assignment, so you can choose a theme that aligns with your group's interests. You can select a problem that you find relevant and exciting to explore. We would like you to use this opportunity to showcase your creativity and develop a data-driven project that you can carry as part of your portfolio of projects.

We encourage you to consider this project an opportunity to develop your programming skills while working on a data-driven project that interests you. This project is an excellent element to highlight your skills and mindset in approaching problems from a data-driven perspective. We hope you enjoy the project and have fun while developing your creativity in relevant data science projects.

Deliveries

1. A short report (in PDF format) with a maximum of 5 pages plus References
2. The PowerPoint presentation slides
3. All Jupyter notebooks (.ipynb) used in the elaboration of the project, please properly comment, and document your code. Please include a text file summarizing the contents of each notebook
4. The data or a representative sample of the datasets that were used for reproducibility
5. The report should include the following elements:

   a. Title;
   b. List of Authors, group number, and class number
   c. 250-word Abstract & three Keywords
   d. Introduction (explain the context of the data and what problem you are trying to study with the data that you are using);

e. Data and Methods (how did you acquire the data? What is the size of the data and its characteristics? Summarize the steps you implemented to clean up the data. Did you use an interesting Python library? Describe it here);
f. Results and Discussion (describe/discuss your main findings and report your results);
g. Conclusions (how does your analysis connect with the problem you propose to study, main challenges, future steps);
h. Statement of Contribution & Acknowledgments
i. References;

Note, if your submission files are too large to submit through Moodle, submit a single PDF with links to download each deliverable separately and a short description.

Grading

The project grading element is worth 60% of your final grade. The grading of the project is divided into the following categories according to the deliverables:

1. Report
    a. Clarity and conciseness: choosing what is relevant to report and what is not; ability to correctly use citations and references.
    b. Clear discussion of data sources and pre-processing steps.
    c. Quality of the discussion and presentation of the methodology.
    d. Ability to motivate each step of the project.
    e. Number of visual aids used (e.g., graphs and charts).
2. Presentation
    a. Clarity and effectiveness of the oral presentation.
    b. Quality and relevance of visual aids (e.g., graphs and charts).
    c. Ability to explain the problem, methodology, and findings compellingly.
    d. Audience engagement and ability to answer questions.
3. Code
    a. Clarity and organization of the code.
    b. Appropriate use of comments and documentation.
    c. Correctness of the code to facilitate data pipeline processing, analysis, and code reuse.
    d. Effective use and choice of libraries and tools given the project goals and tasks.
    e. Inclusion of all Jupyter notebooks used in the project and the ability to reproduce your findings.
4. Deliverables
    a. Inclusion of all required deliverables (report, presentation slides, Jupyter notebooks, and data or representative sample)
    b. Quality and completeness of deliverables

Important Dates

Delivery of a project proposal by May 2nd

Each group needs to submit a short proposal that briefly summarizes the intended project that the group will undertake. Share a title, names and IDs of group members, and a short abstract (250 words max). This will be done in Moodle as a reply to a topic we will create. Please also include a plan for data acquisition (where are you going to collect the data, and what challenges do you foresee)

Final Project Delivery by May 24th

Oral Presentation by May 27th and 28th (1st round) or June 3rd and June 4th (2nd Round)

Prepare a 10-minute presentation plus 5 minutes for discussion. All group members should be prepared to explain a different project section and participate in the debate. We will do the presentations in a blended format.

Previous years

In previous years, popular and exciting projects explored the following lines:

- **Web-scrapping** of data using Libraries such as Scrapy or Beautiful Soup. These projects emphasize the acquisition, cleaning, and pre-processing of data. Examples include:
  - comparing listings on uniplaces.com to compare prices of different offers to find the fair price for a particular offer.
  - Studying the popularity of different beers in different countries from untappd.com.
  - Using data obtained from the portal landing.jobs, a group explored which skills offered a higher salary premium in the IT job market
  - using data collected from football-data.co.uk, a group explored several myths about football: Are there only three significant teams in Portugal? Was Fc Porto the most undisciplined team? Was Sporting CP not playing better than in previous seasons?
- **Lx DataLab** is a data portal maintained by the Municipality of Lisbon that releases challenges that focus on the Municipality's problems. Along with the challenges are datasets shared by their services. These projects are more extensive in scope. In previous years students have used the opportunity to explore the available datasets and report on problems found with the data, inconsistencies, and describe the phenomena. Past challenges embraced by students include a characterization of bike-sharing usage or an understanding of the determinants of road accidents.
- **Explore available structured Open-Data** to tell a story. For instance, in the past, a group used publicly available data (e.g., dados.gov.pt and pordata.pt) to study how different regional factors might inflate or deflate real estate prices in Portugal. On a larger scale, by combining data from the world bank indicators database and the Food Balance dataset, a group explored how differences in food habits between countries could be linked with higher/lower life expectancy. Finally, using data from centraldedados.pt a group studied the recurrence of forest fires in Portugal between 2010 and 2015 and explored possible underlying determinants.
- **Dashboarding**, several groups decided to focus their project on developing an interactive dashboard for storytelling. Often combining multiple datasets, students explored https://plotly.com/dash/ as a framework to create a visualization-driven web portal for storytelling. Naturally, underlying this project was much work to collect and prepare data and test different visualizations.

Tips and Recommendations

- **Start by identifying a question/problem that resonates with all group members**. For instance, do you care about real estate? Or do all of you enjoy Football and its analytical aspects?

- **Finding an interesting question/problem rather than a dataset to work on tends to lead to more interesting, engaging, and fun projects**.
- **List your expected outcomes ahead of starting work on the project.** For instance, visualize a dataset geo-spatially or characterize the relationship between two variables. In other words, identify the elements that will help you build a story and for which there are clear reporting outcomes.
- **From the beginning, be clear about each member's role**. Some will focus more on data acquisition, others on data exploration, and perhaps others on reporting results.
- **We are interested in something other than what works but in the work you have done to develop your story.** For instance, you might list as an expected goal to report a strong positive correlation between two variables. What if you show that there is no correlation at all? It is still interesting; don't throw it away. Adequately tell the story of your de facto results.
- Focusing on working with dirty datasets that require cleaning and preparation is worthwhile, even if you will need to trade off complexity in the project's outcomes. **Data wrangling will also help you develop your programming skills**.
- **Avoid Kaggle projects**. They are boring and have been designed to emphasize modeling performance.
- If you work with structured/cleaned datasets, consider following a project that requires combining multiple datasets.
- **Do not use datasets from contexts that are profoundly irrelevant to us**. For instance, you might find an excellent project on Kaggle about bike sharing in New York, but who cares about New York? Try to find similar datasets for Portugal.
- You are expected to resort to techniques and libraries not necessarily used in the classroom. One of the emphases we put in this curricular unit is to help you develop the ability to independently identify and use the resources you need for each project. For instance, we will not cover data scrapping libraries, but students have often used the project as a context to develop their skills in that sense.
- It is normal to have questions, but we are here to support and help with your project development. Reach out to us if you need help, advice, or support!