Programming for Data Science 2022

Guidelines

As a Data Scientist, you will often be challenged to bring new and valuable insights to problems or questions relevant to your organization or society. More often than not, you will rely on data to obtain such insights. Tasks that you might find yourself doing include: acquisition, filtering, cleaning, and merging of data from multiple sources; classifying information; clustering or segmentation; prepare regression models and/or training machine learning algorithms; identifying non-intuitive relationships between features; test your findings; designing meaningful visualizations to communicate your findings; and more importantly construct a narrative to report your results. Indeed, data science is about building a data-driven compelling story about the problem you have on hands.

The project's goal is to explore a problem of your interest, using python as the main tool of this data science project. It is an open topic assignment; therefore, you are free to choose a theme according to the group interests.

It is expected that the specific steps taken in the elaboration of the project will be different from group to group. Moreover, since this is your first semester in the post-graduation program, we expect projects to focus on data acquisition, data cleaning, and description/exploration of phenomena. Hence, modeling aspects — such as the complexity of the model or its performance — are not going to be considered for the grading.

More importantly, the goal of the project is for you to have fun developing your programming skills while working on a data-driven project. Consider this project as an opportunity to show and advertise your creativity in developing relevant data science projects. You will carry the final project with you as part of your portfolio of projects, which constitutes an excellent element to highlight your skills and your mindset in approaching problems from a data-driven perspective.

Deliveries

1. A short report (in PDF format) with a maximum of 5 pages plus References
2. A PowerPoint with the presentation slides that follow the template shared n Moodle
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. A representative sample of the dataset that was 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 are you 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 connects with the problem you propose to study, main challenges, future steps);
h. Statement of Contribution & Acknowledgments
i. References;

Important Dates

Delivery of a project proposal by April 29th

The project proposal needs to be submitted as a reply to a Topic on the Moodle Discussion Forum. It must include a temporary Title for the project, a short description of the project (goals, context, data sources, expected goals), list of members of the group.

Final Project Delivery by May 22nd

Oral Presentation by May 25th and 26th (1st round) or June 1st and 2nd (2nd Round)

Prepare a 15-minute presentation plus 5 minutes for discussion. All group members should be prepared to explain a different section of the project and participate in the debate. We will try to 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 in 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 the bike-sharing usage or understanding of the determinants of road accidents.

- **Explore available structured Open-Data** to tell a story. For instance, in the past, a group used available publicly available data (e.g., dados.gov.pt and pordata.pt) to study how different regional factors might inflate or not 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** last year, 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 a lot of 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 first 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 to work on the project. For instance, you might want to do a geo-spatial visualization of a dataset or characterize the relationship between two variables. In other words, identify the elements that will help you build up a story and for which there are clear reporting outcomes.
- Be clear, from the beginning, what will be the role of each member. Some will focus more on the data acquisition, others in the data exploration, and perhaps others in the reporting of results.
- We are not interested in what works, but in the work that you have done to develop your story. For instance, you might list as an expected goal to report a positive strong correlation between two variables. What if you show that there is no correlation at all? It is still interesting, don't throw it away, adequate the story telling your de facto results.
- Focusing on working with dirty datasets that require cleaning and preparation is worthwhile, even if for that you will need to trade off complexity in the outcomes of the project. It is in the data wrangling that you will also develop more your programming skills.
- Avoid Kaggle projects. They are boring and have been designed to emphasize modeling performance.
- If you choose to work with structured/cleaned datasets, consider following a project that aims at 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.
- It is expected that you resort to techniques and libraries that have not been necessarily used in the classroom. One of the emphases we put in this curricular unit is in helping 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, we are here to support and help your project development. Reach out to us if you need help, advice, or support!