Final project

KHOA CÔNG NGHỆ THÔNG TIN
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

fit@hcmus

Overview

□ Find a public data (e.g. data on Kaggle) about a subject your group is interested in

□ Explore data (*often interleaved with preprocessing data*)

□ Identify meaningful questions which can be answered with this data

□ Preprocess and analyze data to answer each question

FINAL PROJECT — THINGS NEED TO BE PRESENTED IN YOUR JUPYTER NOTEBOOK

1. Collecting data

- □ What subject is your data about? What is the source of your data?
- □ Do authors of this data allow you to use like this? You can check the data license
- □ How did authors collect data?

2. Exploring data (often interleaved with preprocessing)

- □ How many rows and how many columns?
- □ What is the meaning of each row?
- □ Are there duplicated rows?
- □ What is the meaning of each column?
- □ What is the current data type of each column? Are there columns having inappropriate data types?
- □ With each numerical column, how are values distributed?
  - ◘ What is the percentage of missing values?
  - ◘ Min? max? Are they abnormal?
- □ With each categorical column, how are values distributed?
  - ◘ What is the percentage of missing values?
  - ◘ How many different values? Show a few
  - ◘ Are they abnormal?

3. Asking meaningful questions

- □ Your group needs to give ≥ the-number-of-group-members questions which can be answered with this data.
- □ Each question should be meaningful (what are benefits of finding the answer?) and not too easy to answer (e.g., it’s too easy if we just need one line of code to get the answer).
- □ Your group should focus more on the quality of questions than the quantity
- □ In notebook file, with each question, your group needs to present:
  - □ What is the question?
  - □ What are benefits of finding the answer?

4. Preprocessing + analyzing data to answer each question

- With each question:
  - Does it need to have preprocessing step, and if yes, how does your group preprocess?
    - **Text**: sketch steps clearly so that readers can understand how your group preprocesses even without reading code
    - **Code**: implement sketched steps. Your group should also try to write code clearly (choose good variable names, comment where should be commented, don’t let a line too long)
  - How does your group analyze data to answer the question?
    - Text: similar to above
    - Code: similar to above

5. Reflection

- □ **Each member**: What difficulties have you encountered?
- □ **Each member**: What have you learned?
- □ **Your group**: If you had more time, what would you do?

6. References

□ To finish this project, what materials have you consulted?

Some general points:

- □ In notebook file, one important thing your group should try to practice is to organize sections and write/code clearly. Your group should use Markdown headings to organize sections, and use Jupyter Notebook/Lab TOC to quickly navigate through headings (it’s similar to bookmark in pdf file). During the process of writing/coding, try to maintain a calm mind, try to think for readers
- □ In data science process your group needs to do, I think the most time-consuming part is from finding data to giving meaningful questions which can be answered with data

Final project — Teamwork

- Your group will use Git and Github to do version control as well as to collaborate with each other
- Your group needs to make sure that:
  - The amount of work is quite balance between members (commit history in Github should show that)
  - Members must understand work of each other

Final project — Teamwork

- To achieve this teamwork requirement, with each step in data science process:
    - Each group member will do this step independently → commit history in Github: each member will have his/her own branch and do in this branch (before this, one member will write notebook skeleton and push to Github repo for all members)
    - Then, your group will have a group meeting to understand notebook of each other and produce the group version → commit history in Github: all member branches are merged to main branch
- With the step of preprocessing + analyzing data to answer each question, members can do different questions, but your group needs to make sure members understand work of each other

Final project — Teamwork

- Project grades will be the same for all group members and will be approximately the average of individual member grades; if one member does little and/or does not understand the group notebook well, then grades for all members will be pulled down
- Each group member needs to think for the whole group:
  - “Strong” members should try to support (support ≠ do) “weak” members
  - “Weak” members should try to not pull “strong” members down

Final project — Teamwork

- The process of doing this project can be divided into some phases; before each phase, your group should have a meeting and write down clearly the plan about what tasks need to be done, what amount of time for each task, who will do which (your group can use Google Sheets, Trello, …)
- Your group also needs to write this link (Google Sheets, Trello, …) to README.md file in Github

Final project — Milestones

- x = presentation day (x may be …)
  - x-2, before 9:00 am: each group will upload its project to Github, and open an Github issue to announce the completion state of the project and invite others to review
  - x-2, after 9:00 am: each member of each group will review another group’s project (using Github issue)
  - x-1: each group will revise its project based on reviews, upload the final version to Github and Moodle

Final project — Milestones

- □ x: action! (online)
  - ▣ Each group will have ~15 minutes to present (I will decide who will present which) and ~5 minutes to Q&A; you can present directly on jupyter notebook file, no need to prepare additional slides

  → Your group needs to practice beforehand to make sure that your group can present the notebook with ~15 minutes (when presenting, you should focus on ideas, minimize talking about details and code)