Wrap-up

Marco Morales
mam2519@columbia.edu

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2018
Columbia University

What should be in your presentation?

- keep it **simple**, **concise**
- be prepared to **answer all questions**
- slide limit: **5-8 slides**
- time limit: **5-10 mins**
- **presentation structure**:
  1. **question** you seek to answer
  2. **data** you used
  3. **method(s)** you employed
  4. **insight** (concise, well-defined, actionable...)
  5. your **next steps** for the project
  6. a lenghty **appendix** (that hopefully no one will have to see)

What should be in your presentation?

- make ony **one point per slide**
- rely on **images** as much as possible
- be mindful of **slide headers**
  - slide **headers as one-liners**
  - headers should **summarize the point in the slide**
  - some who only read your headers should know exactly what the presentation was about
- foresee questions and their answers in appendix slides

What should be in your GitHub repo?

- a **well-structured project**:
  - project description (on your repo's landing page)
  - code section
  - data section
  - outputs section
- a **detailed backlog** of what you've done

What should be in your GitHub repo?

a structured project sample

```
project\
|
| -- src
|      |-- data             <- Code to read/munge raw data.
|      |-- features         <- Code to transform/append data.
|      |-- models           <- Code to analyze the data.
|      |-- visualizations   <- Code to generate visualizations.
|
| -- data
|      |-- raw              <- The original, immutable data dump.
|      |-- external         <- Data from third party sources.
|      |-- interim          <- Intermediate transformed data.
|      |-- processed        <- Final processed data set.
|
| -- reports
|      |-- documents        <- Documents synthesizing the analysis.
|      |-- figures          <- Images generated by the code.
|
| -- references             <- Data dictionaries, explanatory materials.
|
| -- README.md              <- High-level project description.
| -- TODO                   <- Future improvements, bug fixes (opt)
| -- LabNotebook            <- Chronological records of project (opt)
```

Sources: **Cookiecutter for Data Science**, **ProjectTemplate**

What should be in your GitHub repo?

An appropriate Project Description

- your project description should summarize all elements present in your project

(1) **Project description**

- What is your project about?
- Keep it simple, narrow and focused

(2) **Insight**

- What is the objective you sought to accomplish?
- Follow the answer to the previous question with a: “so what"? If that’s the end of the conversation, then you need to refine your insight.
- Keep it concise, well-defined and, above all, actionable.

What should be in your GitHub repo?

An appropriate Project Description

(3) **Strategy**

- How did you accomplish your objective?
- Provide a brief description of the method

(4) **Data**

- What data did you leverage to answer your question?
- What are the advantages of that data?
- What are the shortcomings of that data?

(5) **Output**

- Define the parameters that determine that your project is done, your insight delivered, and your output completed.

What should be in your GitHub repo?

- of particular importance: **reproducibility / portability**
- your project **must have**
  - **code and outputs** for
    1. data exploration
    2. data analysis
    3. modeling
    4. visualizations
  - a **summary report** that details the project's insights
- above all **any person should be able to pick up your project and run it / build on it seamlessly**