Programming and Data Science for User Experience Researchers

HCDE 598 A — Jonathan T. Morgan & Ray Hong

Week 9 | May 23, 2016

Overview of the day

- Administrativa
- Writing functions #3
- *10 minute break*
- Final presentation and report - overview and grading
- Course retrospective
- Open space: Review key concepts
- *10 minute break*
- Final project work/Q&A time

Final project presentation

Overview

- Due Friday, June 3rd by 6pm via Canvas (upload slides or link)
- No longer than 5 minutes
- 10% of your grade (30 points)
    - -10 points for every day late

If you are unable to attend on 6/3:

- Let Jonathan know ASAP (no later than Wednesday, 5/25)
- Due Wednesday, June 1st by 6pm via Canvas (upload slides, video file or link)

If you need to attend remotely on 6/3:

- Let Jonathan know ASAP (no later than Wednesday, 5/25)

http://wiki.communitydata.cc/DS4UX_(Spring_2016)#Final_Project_Presentation

Grading rubric - framing

(5 pts) Clearly stated research question or hypothesis

- RQ:
  - "*Which neighborhoods in Seattle have seen the greatest increase in new apartment construction?*"
- Hypothesis:
  - "*The neighborhoods in Seattle that have experienced the greatest increase in new apartment construction will also experience the greatest in increase in demolitions of existing structures*"

(5 pts) Clear motivation for the study: why is this research interesting, relevant, and important?

Grading rubric - data and methods

(5pts) Data is clearly described

- Source(s)
- Type(s)
- granularity

(5pts) Methods are clearly described

- How you gathered your data
- How you performed your analysis

Grading rubric - findings

(5pts) At least 1 finding from the study is presented and described

(5pts) Finding is discussed (pick at least 2 of these)

- How do you interpret these results?
- Are these the results you expected? Why or why not?
- Are there implications for the design of systems?
- Are there implications for the design of policy?
- What kind of further research would be useful/interesting, based on these findings?

Final project report

Overview

- Due Wednesday, June 8th by 6pm via Canvas (upload all files)
  - Code and data in .ZIP folder
  - Report in PDF or Jupyter notebook (link to notebook)
- No more than 3000 words

“A successful project will tell a compelling story and will engage with, and improve upon, the course material to teach an audience that includes me, your classmates, and HCDE students taking classes like this in future years how to take advantage of community data science more effectively.”

http://wiki.communitydata.cc/DS4UX_(Spring_2016)#Final_Project_Report

Grading rubric - components

(10 pts) Code

- Can be executed without errors
- Works with the raw data files you submitted

(10 pts) Raw data and process files

- All raw data...
  - Can be opened in a plain text editor (JSON, TSV, CSV, etc.)
  - Can be processed with included code files.
  - If too big to upload, truncated version is provided
- All process files...
  - Can be generated by running included code files OR
  - For graphs, all pre-processing must be done in Python, not Excel. Include spreadsheet, not just screenshot

Grading rubric - formatting

(10 pts) Basic formatting

- Has cover page OR masthead with title, name, date, course info
- Spacing is 1.5, body font is 10-12 point
- Spelling/grammar checked
- Headings, subheadings, page numbers used consistently and correctly
- Numbers, labels, captions and, citations for all images, graphs, and charts

(10 pts) References

- Has references section
- Citations for all online data sources (URLs for API endpoints and/or documentation, dataset pages… anything with a URL!)
- Citations for all references to previous research studies, news articles, etc
- Has In-text citations for all listed references

Grading rubric - content pt. 1

(10 pts) Introduction

- appropriate background information to motivate the research questions
- All research questions and/or hypotheses clearly stated

(10 pts) Data sources

- All data sources described in sufficient detail, w/ links to online sources
- Examples of data provided (screenshots, table, schema, etc)

(10 pts) Methods

- Data collection methods described in sufficient detail
- Data processing steps described in sufficient detail, w/ links to all code and process files

Grading rubric - content pt. 2

(10 pts) Clearly describe your findings

- if multiple research questions, findings are separated out according to question
- if hypothesis testing, describe whether the hypothesis was supported or not

(10 pts) Discussion and conclusion

- Address at least 2 of the following discussion points
  - how do you interpret these findings?
  - is it what you expected? why or why not?
  - does it have implications for design? (of systems or of policy?)
  - does it suggest further research?
- Limitations and Conclusion
  - Describe at least 1 limitations of study
  - Succinctly summarize your research questions and primary findings

Course retrospective

Why teach programming this way?

Why teach programming this way?

Most programming classes

- **Fundamentals first, applications later (if at all)**
- **Teach programming as a vocation or a science**
- **Focus on software development**

This programming class

- Fundamentals and applications in parallel
- Teaches programming as tool or a type of literacy
- Focus on data manipulation

Why call what we’re doing ‘data science’?

“Science is becoming a database query problem”

- Bill Howe, Director of UW Data Science MS Program

“Science is becoming a database query problem”

Identifying, exploring, hypothesizing about, analyzing, understanding, validating… data

“answering even a simple scientific question requires lots of choices that can shape the results.”

- Christie Aschwanden, “Science Isn’t Broken”
FiveThirtyEight

“80% of analytics is sums and averages”

- Aaron Kimball, CTO Zymergen Inc.

What we covered this quarter

What we covered

1. How to get the data you want from the web
2. How to read complex structured data into a program
3. How to parse, filter, search, separate, and combine data within a program
4. How to write complex structured data out of a program

What we covered

1. How to request and filter data from the web
2. How to read complex structured data into a program
3. How to parse, filter, search, separate, and combine data within a program
4. How to write complex structured data out of a program
5. **How to troubleshoot our own work**
    a. **identifying problems, finding and evaluating possible solutions**

What we covered

1. How to frame research questions that web data can address
2. How to identify and describe datasets that can be used to answer those questions
3. How to explore, validate, and clean your data
4. How to combine data from different sources
5. How to identify the best tool for the job (data structures, loops, other programs)
6. How to analyze your data with those tools
7. How to show your process so that others can evaluate, sanity-check, and **reproduce** your findings

What worked for you?
What didn’t?

Other questions or comments?