Data Management and Life Cycle

(PA397C ı INF385T ı EC0395M)

Course Description

- Instructor: Prof. Ji Ma
- Time: Thursday, 2-5pm, 2023 fall
- Location: SRH 3.316 (in-person only).
- Office hour: Fridays, 2-4pm (pls book to avoid conflict with others).

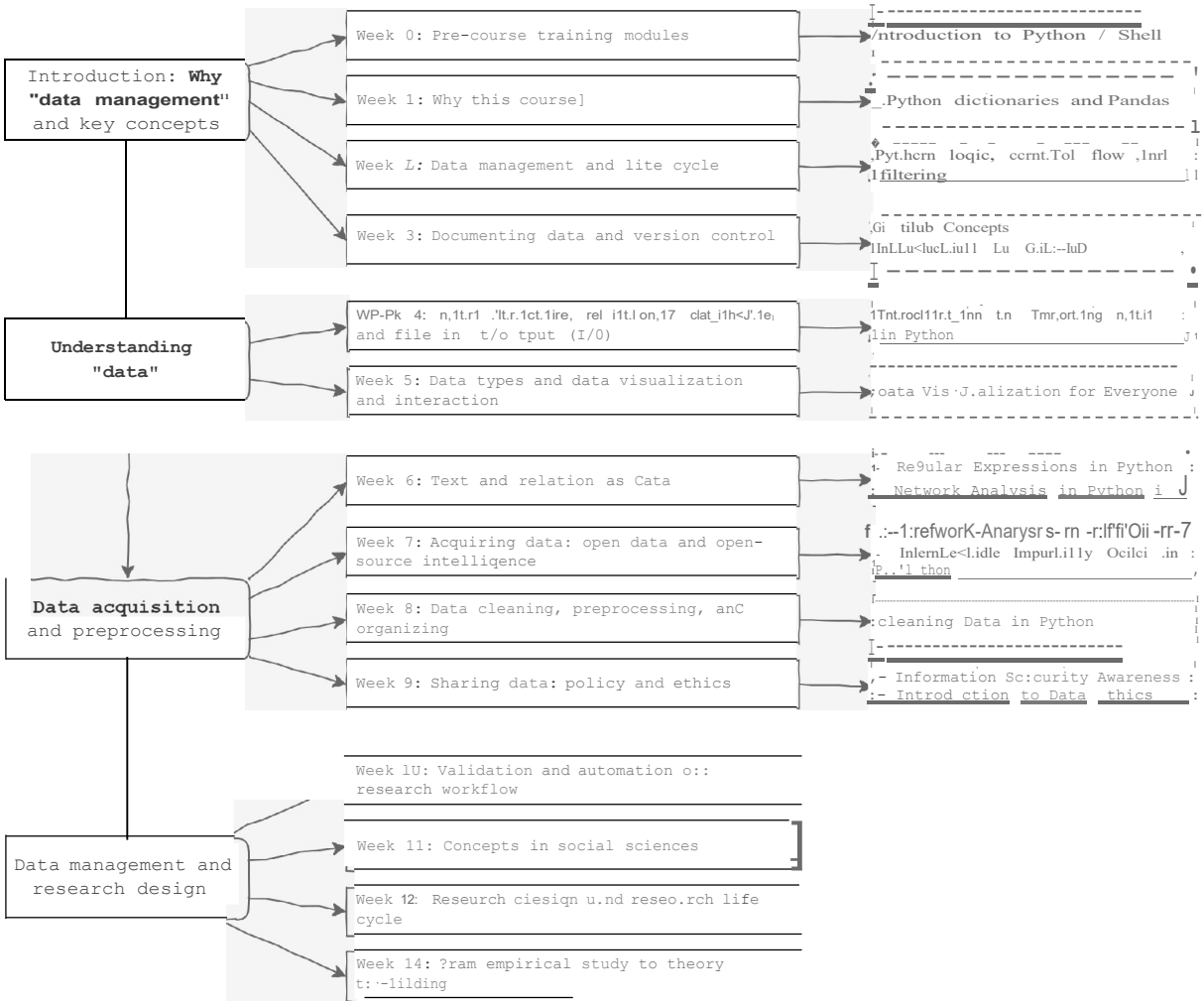
This class equips you with powerful data management skills. You will learn how to manage and work with big, complex, and unstructured datasets in the public and nonprofit sectors. You are expected to learn the following skills and respond to "big questions'' that have social importance: 1) Understand the structure of data and how to work with big and complex datasets; 2) Understand the workflows of acquiring and managing data; 3) Able to conduct data-intensive and replicable social science research.

Programming is not a prerequisite of this class, and you will have a chance to develop your own programming skill set. I primarily uses Python for data work and Stata for statistical analysis, but you are welcome to use any programming language or software as long as you can complete the assignments.

Course roadmap

Data Management and Research Life Cycle

Ji **Ma@UT Austin**



Recommended online tutorials

As a student of this course, you have free access to DataCamp.

Grading

See a list of Assignments

- A>= 98%, A- >= 90
- B+ >= 87%, B >= 83%, B- >= 80%
- C+ >= 77%, C >= 73%, **C-** >= 70%
- D+ >= 67%, D >= 63%, D- >= 60%

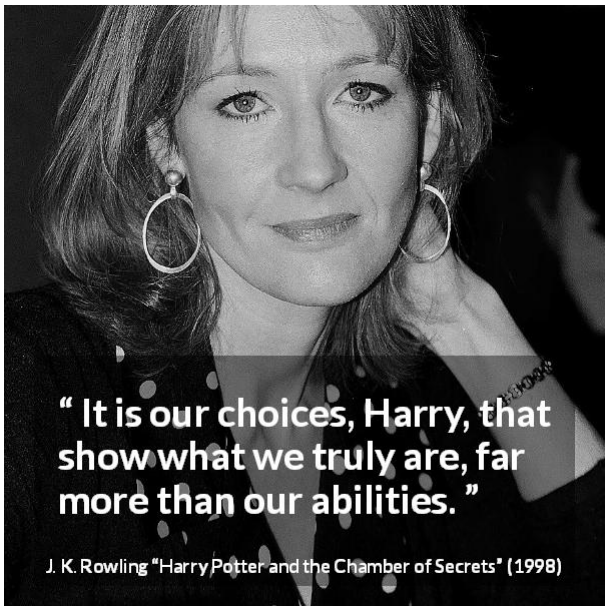
Resources

- Past presentations: 2019 Spring 2020 Spring 2021 Spring

Copyright for Open Education

Data Management and Life Cycle
(PA397C | INF385T | EC0395M)

Course Schedule



- Reading materials by week
- Annotation log and presentation sign-up

The schedule is tentative, we may arrange a few visits to some organizations. Details TBD.

- Week 0 Pre-course

Introduction

- Week 1 8/24: Why this course?
- Week 2 8/31: Data management and life cycle
- Week 3 9/7: Documenting data and version control

Understanding Data

- Week 4 9/14: Data structure, relational database, and data dictionary

- Week 5 9/21: Data types and data visualization and interaction

Data Acquisition and Preprocessing

- Week 6 9/28: Acquiring data: open data and open-source intelligence (guest speaker)
- Week 7 10/5: Text and relation as data
- Week 8 10/12: Data cleaning, preprocessing, and organizing+ Data security
- Week 9 10/19: Field visit: Dress for Success
- Week 10 10/26: Standardization and automation

Data Management and Social Science Research

- Week 11 11/2: Concepts and measures in social sciences (guest speaker)
- Week 12 11/9: Data reuse and data governance
- **Week 13 11/16: Final project workday - no class, the week before Thanksgiving**
- Week 14 11/30: From empirical study to theory building. Final project presentation.

Week 0 Pre-course *Back2Top*

- Learning modules
  - Introduction to Python
  - Recommended: Introduction to Shell
- Pre-course survey

Week 1: Why this course? *Back2Top*

Before class

- Readings:
  - Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature News, 533(7604), 452. doi:10.1038/533452a.
  - Briney, K. (2015). The Data Problem. In Data management for researchers: Organize, maintain and share your data for research

success. Research Skills Series (Exeter, England). HOLLIS number:014921191. Exeter, UK: Pelagic Publishing.

- Gentzkow, M., & Shapiro, J. M. (2014). Introduction. In Code and data for the social sciences: A practitioner's guide.
- Lazer, D., Pentland, A., Adamic, L., Aral, S., Barab'asi, A.-L., Brewer, D., ... Alstyne, M. V. (2009). Computational Social Science. Science, 323(5915), 721-723. doi:10.1126/science.1167742.6

In class

- Discussion and lecture on readings.
- Course review: Syllabus, assignments, final project.

After class

- Learning modules: Intermediate Python
  - Dictionaries & Pandas

Week 2: Data management and life cycle *Back2Top*

Before class

- Readings:
  - Briney, K. (2015). Planning for Data Management. In Data management for researchers: Organize, maintain and share your data for research success. Research Skills Series (Exeter, England). HOLLIS number: 014921191. Exeter, UK: Pelagic Publishing.
  - Briney, K. (2015). The Data Lifecycle. In Data management for researchers: Organize, maintain and share your data for research success. Research Skills Series (Exeter, England). HOLLIS number: 014921191. Exeter, UK: Pelagic Publishing.
  - Ruane, J. M. (2016). Designing Ideas: What Do We Want to Know and How Can We Get There? In Introducing Social Research Methods: Essentials for Getting the Edge (pp. 67-92). Chichester, West Sussex, UK ; Hoboken, NJ: John Wiley & Sons Inc.

In class

- Profile for group matching ("Student Profile")
- Discussion and lecture on readings.
- Review final project possibilities.

After class

- Learning modules: Intermediate Python
  - Logic, Control Flow and Filtering
- **Assignment 1: Plagiarism test (10% points)**

Week 3: Documenting data and version control *Back2Top*

Before class

- Readings:
  - Briney, K. (2015). Documentation. In Data management for researchers: Organize, maintain and share your data for research success. Research Skills Series (Exeter, England). HOLLIS number:014921191. Exeter, UK: Pelagic Publishing.
  - Broman, K. W., & Woo, K. **H.** (2017). Data organization in spreadsheets (tech. rep. No. e3183v1). PeerJ Inc. doi:10.7287/peerj.preprints.3183v1.
  - Gentzkow, M., & Shapiro, J. M. (2014). Version Control. In Code and data for the social sciences: A practitioner's guide.

In class

- Discussion and lecture on readings.
- Student presentation.
- Group discussion on client projects.

After class

- Learning modules:
  - GitHub Concepts
  - Introduction to GitHub

Week 4: Data structure, relational database, and data dictionary

Back2Top

Before class

- Readings:
  - Wickham, H. (2014). Tidy data. The Journal of Statistical Software, 59(10). http://www.jstatsoft.org/v59/i10/
  - Normalization of Database
  - Gentzkow, M., & Shapiro, J. M. (2014). Keys. In Code and data for the social sciences: A practitioner's guide.

In class

- Discussion and lecture on readings.
- Group practice: Form to Table (Your annual happy-hour/nightmare: Form 1040)
- Student presentation.
- Group discussion on client projects.

After class

- Learning modules: Introduction to Importing Data in Python
- Further readings:
  - Swaroop C. H. (2013). Data Structures. In A Byte of Python.

Week 5: Data types and data visualization and interaction

Back2Top

Before class

- Readings:
  - Kirk, A. (2019). Working With Data. In Data Visualisation: A Handbook for Data Driven Design (2nd edition, pp. 95-117). SAGE Publications Ltd.
  - Kirk, A. (2019). The Visualisation Design Process. In Data Visualisation: A Handbook for Data Driven Design (2nd edition, pp. 31-58). SAGE Publications Ltd.

In class

- Discussion and lecture on readings.
- Student presentation: Hajiyeva.
- Group discussion on client projects.

After class

- **Assignment 3 [1/2]: Customized learning - Planned chapters (3% points)**
- **Assignment 5 [1/3]: Client project - Project contract draft (5%)**
- Data Visualization for Everyone

Week 6: Acquiring data: open data and open-source intelligence

Back2Top

Before class:

- Readings
  - Williams, H. J., & Blum, I. (2018). Defining Second Generation Open Source Intelligence (OSINT) for the Defense Enterprise. RAND Corporation.
- Review Bing News Search API: what can you do with it?

In class:

- Discussion and lecture on readings.
- Open dataset/ portal examples
- Group discussion on client projects.
- Hands-on: Bing News Search API (if time allows).

After class

- **Assignment 5 [2/3]: Client project - Project contract final (5%)**
- Learning modules:
  - Network Analysis in Python (Part 2)
  - Intermediate Importing Data in Python

Week 7: Text and relation as data *Back2Top*

Before class

- Readings:
  - Grimmer, J., &Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267-297. doi:10.1093/pan/mps028.
  - Provan, K. G., Veazie, M. A., Staten, L. K., & Teufel-Shone, N. I. (2005). The use of network analysis to strengthen community partnerships. Public Administration Review, 65(5), 603-613.

In class:

- Discussion and lecture on readings.
- Group discussion on client projects.

After class

- Learning modules:
  - Regular Expressions in Python
  - Network Analysis in Python (Part 1)
- Further readings:
  - Borgatti, S. P., & Foster, P. C. (2003). The Network Paradigm in Organizational Research: A Review and Typology. Journal of Management, 29(6), 991-1013. doi:10.1016/S0149-20630300087-4.

Week 8: Data cleaning, preprocessing, and organ1z1ng + Data security *Back2Top*

Before class

Data cleaning, preprocessing, and organizing

- Briney, K. (2015). Organization. In Data management for researchers: Organize, maintain and share your data for research success. Research Skills Series (Exeter, England).

- Gentzkow, M., & Shapiro, J. M. (2014). Directories. In Code and data for the social sciences: A practitioner's guide.
- Miksa, T., Simms, S., Mietchen, D., & Jones, S. (2019). Ten principles for machine-actionable data management plans. PLOS Computational Biology, 15(3), e1006750. doi:10.1371/journal.pcbi.1006750.

Data security

- Briney, K. (2015). Managing sensitive data. In Data management for researchers: Organize, maintain and share your data for research success. Research Skills Series (Exeter, England). HOLLIS number: 014921191. Exeter, UK: Pelagic Publishing.
- Case: UT Data Classification Standard

In class:

- Discussion and lecture on readings.
- Student presentation: Barroso.
- Group discussion on client projects.

After class

Data cleaning, preprocessing, and organizing

- Learning modules:
  - Cleaning Data in Python

Data security

- Learning modules:
  - Information Security Awareness@UTLearn
  - Introduction to Data Ethics

Week 9: Field visit: Dress for Success (3000 S I-35 Frontage Rd Suite 180, Austin, TX 78704) *Back2Top*

Schedule

- 2-3pm:
- 3-4:30pm:

Week 10: Standardization and automation *Back2Top*

Before class

- Required readings:
  - Gentzkow, M., & Shapiro, J. M. (2014). Automation. In Code and data for the social sciences: A practitioner's guide.
  - Wilson, G., Bryan, J., Cranston, K., Kitzes, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. PLOS Computational Biology, 13(6), e1005510. doi:10.1371/journal.pcbi.1005510.
  - ode Visser C, Johansson LF, Kulkarni P, Mei H, Neerincx P, Joeri van der Velde K, et al. (2023) Ten quick tips for building FAIR workflows. PLoS Comput Biol 19(9): e1011369. https://doi.org/10.1371/journal.pcbi.1011369
- Recommended readings:
  - Wilkinson et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. doi:10.1038/sdata.2016.18.
  - Wilson, G., Aruliah, D. A., Brown, C. T., Hong, N. P. C., Davis, M., Guy, R. T., ... Wilson, P. (2014). Best Practices for Scientific Computing. PLOS Biology, 12(1), e1001745. doi:10.1371/journal.pbio.1001745.
  - Gentzkow, M., & Shapiro, J. M. (2014). Appendix: Code Style. In Code and data for the social sciences: A practitioner's guide.

In class

- Discussion and lecture on readings.
- Goodenough practices in final project
- Student presentation: Ramarao, Wang.
- Group discussion on client projects.

After class

Work on your customized learning modules.

Week 11: Concepts and measures 1n social sciences *Back2Top*

Before class

- Required readings:
  - Ruane, J. M. (2016). All That Glitters Is Not Gold: Assessing the Validity and Reliability of Measures. In Introducing Social Research Methods: Essentials for Getting the Edge (pp. 117-138). Chichester, West Sussex, UK ; Hoboken, NJ: John Wiley & Sons Inc.
  - Ruane, J. M. (2016). Measure by Measure: Developing Measures-Making the Abstract Concrete. In Introducing Social Research Methods: Essentials for Getting the Edge (pp. 93-116). Chichester, West Sussex, UK ; Hoboken, NJ: John Wiley & Sons Inc.
  - Shoemaker, P. J., Tankard, J. W., & Lasorsa, D. L. (2003). Theoretical Concepts: The Building Blocks of Theory. In How to Build Social Science Theories (pp. 15-36). SAGE Publications.
- Recommended readings:
  - Gerring, J. (1999). What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences. Polity, 31(3), 357-393. doi:10.2307/3235246.

In class

Guest speaker: Lifelong Learning with Friends (2pm)

For the client (30 mins with Q&A):

- What and how data are generated in daily operations.
- How data are processed, shared, and collaborated in daily operations.
- What are the relations between data and business, and who are the users of the data.

*For the student team** (20 mins with Q&A)*

- Where is the niche for the team to fit in?
- What are the deliverables and how they can be useful?

Weekly class activities

- Discussion and lecture on readings.
- Student presentation: Raza, Chavera.

- Group discussion on client projects.

Week 12: Data reuse and data governance *Back2Top*

Before class

- Readings:
  - Briney, K. (2015). Data reuse and restarting the data lifecycle. In Data management for researchers: Organize, maintain and share your data for research success. Research Skills Series (Exeter, England).
  - Ghavami, P. (2020). Data Governance and Data Security. In Big Data Management: Data Governance Principles for Big Data Analytics. De Gruyter.

In class

- Discussion and lecture on readings.
- Student presentation: Vanegas, Zhang.
- Group discussion on client projects.

After class

Make sure you and your team are on track of all outstanding assignments.

Week 13: Final project workday - no class (week before Thanksgiving) *Back2Top*

Assignment due:

- **Assignment 3 [2/2]: Customized learning - Completion (27% points)**

Also work on any outstanding assignments:

- Analysis of empirical studies and presentation.
- Client project.

Week 14: From empirical study to theory building. Final project presentation *Back2Top*

Before class

- Readings:
  - Creswell, J. W. (2014). The Use of Theory. In Research design: Qualitative, quantitative, and mixed methods approaches (4th ed). Thousand Oaks: SAGE Publications.
  - Sutton, R. I., & Staw, B. M. (1995). What Theory is Not. Administrative Science Quarterly, 40(3), 371-384. doi:10.2307/2393788.
  - Shoemaker, P. J., Tankard, J. W., & Lasorsa, D. L. (2003). Theoretical and Operational Linkages. In How to Build Social Science Theories. SAGE Publications.
  - *(review the article wrote by the authors in Week 7)* Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., Freelon, D., Gonzalez-Bailon, S., King, G., Margetts, H., Nelson, A., Salganik, M. J., Strohmaier, M., Vespignani, A., & Wagner, C. (2020). Computational social science: Obstacles and opportunities. Science, 369(6507), 1060-1062. https://doi.org/10.1126/science.aaz8170

In class

- Discussion and lecture on readings.
- Final project presentations.

After class

- Further readings:
  - Shoemaker, P. J., Tankard, J. W., & Lasorsa, D. L. (2003). Creativity and Theory Building. In How to Build Social Science Theories (pp. 145-166). SAGE Publications.
- **Assignment 5 [3/3]: Client project - Final report and presentation (30%)**
- Also make sure you clear outstanding submissions for these assignments:
  - **Assignment 2: Annotation on weekly readings (5% points)**
  - **Assignment 4: Analysis of empirical studies (15% points, due: week of your selection)**

Data Management and Life Cycle
(PA397C ⁞ INF385T ⁞ EC0395M)

- Assignment 1: Plagiarism test (individual, 10 points)
- Assignment 2: Annotation on weekly readings (individual, 5 points, due: each week before class)
- Assignment 3: Customized learning (individual, 30 points)
- Assignment 4: Analysis of empirical studies (individual, 15 points, due: week of your selection)
- Assignment 5: Client project (group, 40 points)

Due dates are listed on Canvas. Late submissions are not accepted.

Assignment 1: Plagiarism test

The first assignment of this course is to pass the plagiarism test and obtain a certificate at the master and doctoral level. Plagiarism is a serious academic misconduct. You will receive zero grade on plagiarized work and there may be other consequences. We have been told not to do this maybe since primary school, and we are always assuming we know what plagiarism is. However, we may assume we know too much (e.g., Notable Cases of Plagiarism).

You do not need to take this test if you have comparable certification, but the validity of your certification needs to be approved by the instructor.

"All assignments in this course may be processed by Turnltln, a tool that compares submitted material to an archived database of published work to check for potential plagiarism. Other methods may also be used to determine if a paper is the student's original work. Regardless of the results of any Turnltln submission, the faculty member will make the final determination as to whether or not a paper has been plagiarized" (Statement from the Faculty Writing Committee: Guidelines for Preventing Plagiarism).

For this assignment, please submit your certificate as a file.

Assignment 2: Annotation on weekly readings

You are required to read the reading materials before class, at least have one comment on each article, and respond to at least one comment from another classmate. This routine assignment uses online annotation platform, and is **due before class day.**

Log your annotations here for grading.

For this assignment, you only need to log your reading progress, no submission required.

Assignment 3: Customized learning

I have listed relevant DataCamp online modules after each week for your reference. You can also choose the chapters or modules of your interests. Each chapter is worth 3 points, you should complete at least 9 chapters.

This assignment has two sub-assignments:

[1] Planned chapters (3% points): Please submit a list of your planned chapters with expected due dates.

[2] Completion (27% points): Please submit all completion evidence in one single file (e.g., screenshots, certification file, etc.).

Assignment 4: Analysis of empirical studies

- Sign-up sheet for individual presentation
- Past presentations: 2019 Spring 2020 Spring 2021 Spring

You are expected to present an analysis of 2-3 empirical studies published by the top journals in respective field using the knowledge learned from the reading materials. The empirical studies can be research reports, academic journal articles, or any other evidence-based and data-driven studies/projects. The materials of the empirical studies need to be uploaded to the online course folder at least a week before class, and slides (e.g., PowerPoint Slides) for presentation are required (no due date).

Example questions:

- *Research design questions:* What is the research question? Where does the data come from? How does the author(s) operationalize the research question(s) using the data?
- *Workflow and replication questions:* How does the author(s) organize the files? Is the study replicable? How can its reproducibility be improved?
- *Validation questions:* How does the author(s) validate the data? Are the conclusions still valid if different datasets are used?
- *Critiques:* How can the study be improved regarding its data selection, validation, and management? Are there any methodological flaws regarding its use of data? Are there any logical gaps between data and empirical analysis?

Rubric. Your presentation should clearly demonstrate the studies':

- Research question, research design (e.g., how abstract concepts are measured), and their relations with data (30%).
- Data management, documentation, and reproducibility (30%).
- Computational methods used, and how data is analyzed using these methods (overview is adequate; 30%).
- Critiques (10%).
- Keep your presentation within 30 minutes.

For this assignment, please submit your presentation slides as a file via Canvas.

Assignment 5: Client project

- Needs Assessment Questionnaire
- Needs Assessment Submitted by Clients
- Final Project Assessment Scale

Your group are expected to work with a client to help them solve a data management problem. Over the past few years, we've successfully executed a multitude of data projects, enabling dozens of nonprofit organizations and government departments to meet their data management objectives. The presentations of some of these projects can be accessed here: 2019 Spring/ 2020 Spring/ 2021 Spring

Specific deliverables and milestones for this group project:

1. Project contract draft (5 points)

No required format for the contract. It should serve as a concrete plan and guideline for your team and an agreement between your team and client. It usually includes these components: project objective, deliverables, member responsibility, timeline, conditions of confidentiality, communication methods, signatures of team members and client representative, etc.

Submit the draft contract as a file via Canvas.

2. Project contract final (5 points)

Submit the final contract with all parties' signatures via Canvas.

3. Final report and presentation (30 points)

Please submit your final report and presentation slides as files via Canvas.

- Please limit your presentation to 30 minutes.
- Your presentation will be evaluated by the class using this assessment scale.
- The instructor will assess your presentation and final report using the same scale. The final grade for this assignment: students' evaluation 30%, instructor's evaluation 70%.
- You will have a week to revise your final report and related files according the feedback received during presentation.
- Please submit your presentation slides, a narrative final report, and any related files and appendix.
- No required format for the final report. But it should detail all your work in a narrative format with relevant references cited. No required citation format as long as it is consistent. Font size 12, 1-inch margins, single-space, 10 pages maximum. You can add more content using an appendix. No required format and no page limit for the appendix.
- If you have problems with uploading multiple files via Canvas, try to upload the files as a compressed file (e.g., `.zip` file). You can also send me your final files via email (not recommended, but okay). The general principle is to have multiple files in one place.