

Course Syllabus as of August 16, 2017

Instructor:	Dr. Andrea Wiggins (she/her/hers)	E-Mail:	wiggins@unomaha.edu
Office:	PKI 177C	Phone:	402-554-2093
Office hours:	Thu, 3 - 4:30 PM & by appointment	Classes held:	Wed, 1:30 - 4:10 PM, PKI 274
		Class dates:	Aug 23 - Dec 6, 2017

1 Course Description

ISQA 4000/8086 Special Topics “From Data to Decisions” provides a hands-on introduction to contextualized data analysis and information management skills. Topics include locating and evaluating data for decision support; data organization, manipulation, analysis, and interpretation; data and process documentation; and presentation of supporting data and recommendations designed for a specific target audience. Ethical, legal, and social implications of managing and using data to support decisions in diverse organizational contexts are discussed and applied in inquiry-driven analysis using command line, SQL, R Studio, Markdown, and Git.

2 Learning Outcomes

Upon completion of the course, students are expected to be able to:

1. Find and evaluate data for a specific purpose such as decision support
2. Understand the limits of data and the questions that can be asked
3. Demonstrate understanding of the contexts of both production and use of data
4. Perform basic data cleaning and management
5. Document data manipulation & analysis processes for reproducible work
6. Demonstrate basic skills & familiarity with:
 - Command line (bash/shell) for navigation and checking file information,
 - SQL (SELECT statements) for retrieving data,
 - R with RStudio for basic descriptive statistics and plotting,
 - Git (via GitHub) for version control in collaborative projects with Markdown for display.
7. Develop a profile of the target audience for analysis results
8. Present information for decision support tailored for the target audience.

Course assignments will give students the opportunity to develop and refine skills with foundational tools for information management, work collaboratively on data-intensive projects, and encourage discussion of the social and organizational aspects of information management practices.

3 Course policies

3.1 Prerequisites

This course has no prerequisites.

3.2 Course Materials

Text

Articles will be provided through Canvas or your library login.

Systems

Students are required to **bring a laptop to every class session** unless otherwise specified, as we will regularly do in-class tutorials and exercises that require capabilities not available on tablets or smartphones.

Before the second week of class, students should download and install the following software:

- GitHub Desktop
- Microsoft Excel
- R Studio

Both GitHub and R Studio are free and open source software. A full-featured text editor is also recommended, such as TextMate (also free) or TextWrangler. The text editor that comes with your operating system is not full-featured.

Students will also need to create accounts for GitHub and Wikipedia *or* Wikimedia.

Please note: Most course assignments will be hosted on GitHub, which is publicly visible. For most students, this is not a problem, as you can always remove the repository after the course is finished, and you do not have to associate your real name with your user account. Associating your real identity with your repositories can be a benefit for some students, as it provides evidence of your ability to use information management tools and to collaborate effectively on a team. However, if you have IP or privacy concerns, please contact me by September 5 to discuss alternate options.

Assignments will be submitted on Canvas, and will typically be a URL to a document or repository on GitHub. All grades and announcements will be available through Canvas as well. Any course materials such as articles or data sets that are under copyright will be provided through Canvas in the Files section, or as links in the weekly module.

3.3 Academic Integrity

Students are expected to follow the University Code of Conduct and are responsible for upholding these standards in this course. It is important to be aware of the consequences of academic dishonesty. Academic dishonesty includes cheating, falsifying information, facilitating others' cheating, plagiarism, buying assignments, submitting the same paper more than once, forging signatures, bribery, and other acts that deceive others about your academic work or record.

Plagiarism carries multiple penalties, in addition to making me sad. Any plagiarism will incur a 0 score on the assignment and a formal report to the University, which will result in a note about the offense added to your student record, and potential further penalties as imposed by the University, up to and including expulsion. If you have any questions about this policy or how to

properly cite materials, please use all available resources, including the library, websites, and me. **All assignments must reflect your own original work.**

3.4 Attendance & Student Conduct

Regular class attendance is obligatory. Since in-class participation is part of the course evaluation, missing class will negatively affect your course grade. In addition, the course is designed to scaffold key activities and content to support your individual assignments and team deliverables, so more than minimal absences will undermine your ability to perform well in this course.

If you must miss class, please notify me in advance by email, and check with your classmates afterward so that you can catch up. This includes any intended absences for religious observances.

As a graduate student, I expect you are fully capable of behaving professionally in the classroom, which means treating every person who enters our classroom with the respect that you would like to experience yourself. This means that:

- side conversations are discouraged,
- your cell phone must be silenced before the start of class,
- you should be using your electronic devices for class purposes only, and
- disruptive students will be asked to leave and will forfeit the participation grade for the day.

3.5 Communications

Communication outside of class will use Canvas or your unomaha.edu email account. Course announcements will be posted on Canvas and individual correspondence will be conducted via Canvas or email. I will make every effort to send announcements with adequate advance notice; failure to receive email announcements will not be considered a suitable excuse for not being informed. ***Include “ISQA 8086” in email subject lines for prompt response; messages without the course number in the subject line may be overlooked.*** I will typically reply in two business days, usually less, but do not expect replies over the weekend. Telephone is not an effective way to contact me, even in emergencies.

3.6 Disability Accommodations

According to University policy, accommodations will be provided for students with verified disabilities who request accommodations. For further information, contact the UNO Accommodation Services Center.

3.7 Intellectual Property

Assignments in this course require adherence to licensing and terms of use as applied to data, which will be discussed further in class. If you have any concerns about the IP constraints on a data source you wish to use, please ask sooner rather than later.

I may request written permission to use exceptional student work as examples for future classes, but you hold all copyright to your own work and may decide whether or not to permit such use. By extension, *you do not have the right to reuse or redistribute any work of your classmates without their consent*, so you and your teammates should agree on acceptable uses of any team project materials for other purposes, e.g., professional portfolios. Be sure to set your GitHub repository license to the appropriate terms; note that the default licenses available do not include CC-BY, so unless you choose a default license, you must explicitly declare the license terms in your Readme.md.

4 Course Schedule

Note the course schedule is subject to change. Assignment due dates are also posted in Canvas.

Readings are listed for the dates for which they should be completed; for example, you should read the selections listed under Week 2 by class time on August 30, etc. Please note that some readings are available via URL instead of a PDF.

Week 1, August 23: Getting Started

- *Administrivia* Review of syllabus and course policies
- *Topics* Course introduction; intellectual property basics; Git & command line basics
- *Assignment* Skills survey at start of class
- *Activities* Git install & tutorial; command line basics

Week 2, August 30: Information Seeking

- *Topics* Types of data (qual/quant & NOIR); information seeking & information overload
- *Readings* Choo et al. (2000); Hemp (2009); Wilson (2017)
- *Assignment* Readme.md - Individual
- *Activities* Information seeking exercises; Project teams revealed

Week 3, September 6: Turning Data into Information

- *Topics* Asking good questions; where data comes from
- *Readings* Davis (1971); O’Neil (2014); Thorp (2017)
- *Assignment* Information Seeking - Individual
- *Activities* Developing research questions

Week 4, September 13: Organizing Data

- *Topics* Metadata, standards, and interoperability
- *Readings* Dabbish et al. (2012), Corti et al. (2014)
- *Assignment* Data entry analysis - Individual
- *Activities* Data “pitches” and in-class ranking of team project preferences

Week 5, September 20: Cleaning Data I

- *Topics* Data & ethics; practical strategies for data cleaning
- *Readings* Van den Broeck et al. (2005); Vitak et al. (2016)
- *Assignment* Project Work Plan - Team; post RQs - Team
- *Activities* Data cleaning with Excel

Week 6, September 27: Cleaning Data II

- *Topics* More data cleaning strategies; integrating, normalizing, & standardizing data
- *Readings* Viswanathan & Viswanathan (2015) Chapter 1 (skip sections on reading in XML, JSON, fixed-width)
- *Assignment* Peer Review 1
- *Activities* Data cleaning with R

Week 7, October 4: Documenting Data Processing

- *Topics* Process documentation in organizations and teams; intro to BPM
- *Readings* Lethbridge et al. (2003); Moreau et al. (2008); Creately (2014)
- *Activities* Develop documentation standards; process diagrams; mid-semester course feedback

Week 8, October 11: Relational Databases

- *Topics* SQL basics (SELECT) (+JOIN)
- *Readings* Halfaker (2016); Wikimedia (2015); Welser et al. (2011)
- *Assignment* Set up a user account on either Wikipedia or Wikimedia
- *Activities* Querying Quarry

Week 9, October 18: Project Progress Meetings

- *Topics* Team meetings & work session
- *Assignment* Draft data cleaning documentation - Team
- *Activities* Team meetings with instructor

Week 10, October 25: Basic Descriptive Analysis

- *Topics* Basic descriptive statistics
- *Readings* Quick (2009); Viswanathan & Viswanathan (2015) Chapter 2
- *Activities* R tutorial

Week 11, November 1: Basic Data Visualization

- *Topics* Basic plotting in R; interpreting descriptive statistics
- *Readings* Hullman & Diakopoulos (2011); Tufte (1983) Chapter 2
- *Activities* Data interpretation
- *Assignment* Draft R Script - Team

Week 12, November 8: Data to Information Workflows

- *Topics* R Studio as a workflow tool; troubleshooting R
- *Readings* Georgakopoulos et al. (1995); Viswanathan & Viswanathan (2015) Chapter 8
- *Assignment* Draft R Plot - Team
- *Activities* R Studio workflow exercise

Week 13, November 15: Audience Analysis

- *Topics* Telling stories with data
- *Readings* Maskiewicz & Kozar (2011); Gershon & Page (2001)
- *Activities* Audience personas

Week 14, November 22: People's Choice

- *Topics* TBD—we may do some catch-up, troubleshooting, studio critique of visualizations, or discuss other topics of interest to the class
- *Assignment* Post Personas by 10 AM Saturday; Peer Review 2

Week 15, December 6: Final Presentations

- *Assignment* Presentations; Team project repositories
- *Activities* Snacks!

5 Assessment

As shown in the table below, 55% of your grade will come from your team projects, 15% from participation, and 30% from individual assignments. This will give you opportunity to demonstrate your mastery of course concepts both independently and as part of your teams, much as in the professional world.

Notably, participation makes up a substantial portion of your grade. Employers will expect you to speak up and share your insights and expertise, so class discussions are good professional practice.

Type	Assignment	Due	Points	Objectives
Individual	In-class Participation	throughout	15	all
Individual	Readme.md	Week 2	5	6
Individual	Info Seeking	Week 3	10	1, 2, 3, 7
Individual	Data Entry Analysis	Week 4	5	4,5
Individual	Peer Review 1	Week 6	5	1
Individual	Peer Review 2	Week 14	5	7
Team	Work Plan & RQs	Week 5	5	2,3
Team	Data Cleaning	Week 9	5	4,5
Team	Progress Meeting	Week 9	5	all
Team	R Script	Week 11	5	5, 6
Team	R Plot	Week 12	5	5, 6
Team	Audience Persona	Week 14	5	7
Team	Project Presentation	Week 15	10	8
Team	Git Repository	Week 15	15	all
Total			100	

5.1 Guidelines for preparing assignments

If the instructions for an assignment aren't clear, *ask*, don't assume. If you have questions about assignments, ask before they are due; do not count on getting a reply the day the assignment is due.

Prepare a professional document with tables, graphs, and references that support your content where appropriate. *Follow all instructions carefully*, and ask questions as soon as they arise if you are uncertain about the assignment requirements.

Assignments that do not meet specified requirements *will* lose points on the assignment grade. It is essential that you spell check and proofread your documents. For written assignments, proof-reading a *printed* copy of your work is especially effective for finding errors that you might overlook on the screen.

General requirements:

- All Markdown documents must be nicely formatted and easy to read. Use section headers and sub-sections where appropriate. Keep it simple.
- All written content must have a blank line between paragraphs (block format) OR the first line of the paragraph must be indented.

- Your name or your team name must appear at the start of every submitted document.
- References must be included at the end of the document for any cited material (including data sets). Use APA format for citations and references. **Web resources must always include the URL and date accessed** regardless of what you may see in examples.
- Some assignments can be submitted in multiple formats; see assignment details for specifics.

The final commit on Git documents must be timestamped before 1:30 PM on the date the assignment is due; the last commit prior to the cutoff time will be graded. Note that commits made from GitHub for Desktop may not appear immediately in your repo, so if you're working to a deadline, committing a draft to generate a URL to submit on Canvas is a good strategy; as long as the final commit is made prior to the deadline, it will be fine.

5.2 Grades and Grading

Assignments are due as defined in the syllabus unless otherwise specified. The penalty for late assignments will be **10% within the first 24 hours, and an additional 25% for each week thereafter**. An exception is possible in an extreme circumstance in which there is no reasonable way to anticipate or control the situation. Computers crashing, viruses, lost files, etc. are specifically not grounds for an extension.

Grading rubrics for each assignment are provided on Canvas; please take advantage of them as you prepare your assignments to check whether your work meets grading criteria. If you wish to discuss a grade, submit a *written* explanation of your argument (email) and arrange for a private conversation. Except for unusual circumstances, no appeals will be considered more than two weeks after the graded paper is returned. For final course grades, no appeal will be considered more than two months after the final day of classes.

Unless announced otherwise, assignments submitted by the due date will be graded within 1–2 weeks. Assignments submitted late will receive lower priority and will take longer to grade. Final grades will be computed based on the scale below and partial points/percentages will be rounded for final grades. The choice to round up or round down is entirely at the instructor's discretion and is generally determined by your overall performance in the course.

96 - 100%+: A (4.0)
 90 - 95%: A- (3.7)
 87 - 89%: B+ (3.3)
 83 - 86%: B (3.0)
 80 - 82%: B- (2.7)
 75 - 79%: C+ (2.3)
 70 - 74%: C (2.0)
 66 - 69%: D (1.0)
 0 - 65%: F (0.0)

5.3 Individual Assignments

45% total

5.3.1 In-class Participation

15%

Everyone is expected to partake in discussion of readings, presentations, and in-class activities. Non-attendance will be reflected in a decrease in this grade (and likely other grades as well). Attendance alone does not ensure credit, as you are expected to actively engage in course activities and discussions.

5.3.2 Readme.md

5% total

Complete the entire Markdown tutorial at <http://www.markdowntutorial.com>. Make a note of the equipment mentioned in the final exercise.

Compose a “readme” document in Markdown to serve as the front page for your personal repository, where you will submit your Information Seeking assignment. At a minimum, this document must include license information for the repository (the choice of license is up to you), and a list of files in the repository. There will not be any other files until the next assignment, of course, so set up a placeholder that you will update as part of the Information Seeking assignment. Submit the URL to the Readme.md page on Canvas, and list the equipment items mentioned in the final exercise in the comment on the submission.

5.3.3 Information Seeking

10% total

For this assignment, you will identify 3 unique, openly available data sets from different sources and summarize key details. The data sets for the team projects will be selected from this assignment, so set aside plenty of time to search for interesting data. Your goal is to make it easy for me to pick awesome data sets for your teams to use, but without having to look at every single data set myself.

You can select any data you think is adequately interesting, with some caveats. You can ask for clarification if you think you have found a worthwhile exception, but in general, please avoid the following:

- Extensive free text data; other qualitative and mixed data is fine if it’s at least nominal or categorical, such as multiple choice answers.
- Enormous data sets are burdensome for class projects; in most cases, you should select data under 1 GB in volume, though there may be some exceptions.
- Data with restrictions that limit reuse and sharing (we will discuss IP in Week 1; ask questions if you’re not certain).
- Overly simplistic data: 2 columns is too trivial for grad school!
- Super complicated data: 20 tables would be a real pain to organize into usable files, but if the data are truly awesome, it might be worth it.
- Data that contain personally-identifiable information (PII) such as full names, email addresses, contact details, etc. Exception: PII is OK when the persons are public figures, e.g., politicians, movie stars, etc. There may be other exceptions; ask if you’re not certain.
- Data about children or other clearly objectionable or morally questionable content (no R-rated data, please).

In addition, the following sources are *not* permitted: Pew Foundation, US Census, Kaggle, Twitter, or Facebook. However, if you find a data set on Kaggle that you are able to properly source for the *original* data set, it’s fair game.

Note that some of the most interesting data may be available to use by permission but not fully available online, e.g., UNO’s Food Pantry keeps records of donations. If you want to propose one of these data sets and are *certain* that the data can be acquired upon request and used for class

purposes, provide the same information (omitting the URL) and add contact details for acquiring the data.

For each data set, create a Markdown document and provide the following information in the following order:

1. The name of the data set as a header
2. About one paragraph describing why these data are interesting
3. A description of how much data there are, in what formats, etc. This should include how many rows & columns (&/or tables) and sample column headers.
4. Details of the license or terms of use (include a link if needed)
5. Potential data users and decision-makers for this data
6. Three questions this data might help to answer; note additional data sources that would be needed, if applicable
7. An APA-formatted data citation (including URL)

Commit the files to your GitHub repository, update the repository Readme.md in keeping, and submit the repository URL on Canvas.

A few students will be asked to “pitch” a data set that they have identified. If you are one of the lucky contestants, you will be notified by email in advance and should prepare a brief pitch (no slides) to introduce the class to the data for consideration for team projects.

5.3.4 Data Entry Analysis **5% total**

For this assignment, you will use the provided files that will be linked to weekly module and the assignment, and answer specific questions about structuring data files for usability.

5.3.5 Peer Reviews 1 & 2 **5% & 5%**

Each student will make constructive replies on three other teams’ research questions and personas in the Canvas Discussion forums during Week 6 and Week 14. The replies do not need to be lengthy, but should demonstrate a thoughtful attempt to provide useful feedback that can help the teams refine their research questions and personas. Clarifying questions, suggestions on wording, and related data sources or information that might inform the work are a few examples of useful contributions to the discussion. In addition, teams should monitor replies to their posts and provide clarifications if questions arise to help classmates provide more useful feedback.

5.4 Team Project **55% total**

Your team project will involve an end-to-end workflow that moves from data discovery through cleaning, documentation, basic analysis, and presentation to a target audience. The goal of the project is to take “raw” data and use it to generate a presentation intended to influence the decisions of specific individuals, while providing full documentation of your process. This is very similar to many real-world team projects, and even those students who are familiar with some of the tools or steps in this process will find something to learn from the experience of developing a fully documented start-to-finish data-driven decision support deliverable. In the process, you will also be introduced to a set of tools that are commonly used in data science and information management more broadly. All written components should be thoroughly proofed for grammar and spelling.

The process is strongly scaffolded to help you avoid procrastinating and get just-in-time feedback; you are expected to take all feedback into consideration in developing your final project repository and presentation. Most of the interim deliverables are “lightly graded” which means that as long as the required items are included and in decent shape, most teams will get full credit for team deliverables up to the final presentation and Git package.

Here’s the hardest part: this is a choose-your-own-adventure and the sophistication of the analysis is up to you! Challenge yourself just enough to learn something new and fulfill the spirit of the assignment, but don’t let it drag you under; you will get plenty of feedback along the way to help you stay on track. If you need additional assistance beyond in-class and assignment feedback, please drop in during office hours or make an appointment.

This project will include:

1. Work plan & draft research questions: Week 6, 5 points
2. Data cleaning documentation draft: Week 9, 5 points
3. Team progress meeting: Week 9, 5 points
4. R script draft: Week 11, 5 points
5. R plot draft: Week 12, 5 points
6. Audience persona: Week 14, 5 points
7. Presentation: Week 15, 10 points
8. Git package: Week 15, 15 points

I will identify viable data sets from your Information Seeking assignments, and they will be assigned to teams based on your preferences. Team membership will also be assigned to ensure a fair distribution of pre-existing skill sets.

5.4.1 Work Plan & Draft Research Questions 5%

Your work plan should specify your draft research questions, planned timeline for completing the project, projected effort allocation (who will do which tasks), and target audience for the analysis. It is an informal document intended to provide background for the progress meeting with the instructor. The work plan should be a simple document (in Markdown) containing the items mentioned above and made available from your team’s Git repository. There are two steps to submit this deliverable: submit the URL for the work plan on Canvas, and then copy out the basic description of your data set and your research questions into a reply on the Discussion thread on Canvas for peer review.

You can earn 1 point of extra credit if your work plan includes a Gantt chart or other diagram to help clarify your plans.

5.4.2 Data Cleaning & Documentation 5%

Your data cleaning and documentation draft will be part of your final project repository. Take advantage of the opportunity by submitting as complete a document as you can muster! If you are integrating multiple data sources, provide (shorter) background details for each data set.

At a minimum, the documentation must include:

1. 1–2 paragraph text description of the data source/s (how much, where from, what it contains, etc.) with a properly formatted citation for each data source. This should include how many rows & columns (&/or tables) and sample column headers.
2. Specifically identify any intellectual policy constraints, or lack thereof (licensing).
3. 1 paragraph description of the metadata: what information is available to help you interpret and understand the data?
4. Identify any issues you have encountered with the data: missing values, unstandardized content, entity matching, etc.
5. 1 paragraph description of your rationale for the steps you’re taking to remediate data. For example, if you need to fill in empty fields, specify what value you chose and why.
6. A script or step-by-step textual description (or a combination) that documents your data cleaning process with enough detail for replication.

This deliverable supports timely feedback for work-in-progress. Since most of you will use data that is much “cleaner” than you would normally encounter in the wild, incomplete data processing documents are acceptable *only* if you can clearly identify the barriers (or series of barriers) to completion, which will help me help you troubleshoot. Any issues highlighted by instructor feedback should be carefully addressed in your final project data processing documentation. This document can be either a descriptive Markdown document with embedded code snippets, or an R script; choose the best one for your project needs and submit the URL on Canvas.

5.4.3 Progress Meeting 5%

The team progress meeting with the instructor must be attended (in person or virtually) by all team members for full credit. This is an excellent opportunity to resolve questions, ask for advice on how to move forward effectively, and verify that you’re pursuing a viable direction for your team project. The meeting will be based on the Data Cleaning & Documentation assignment you submit prior to class; be prepared to provide a quick rundown of your plans and main concerns or questions. The team progress meetings will be held during class time.

5.4.4 R Script Draft 5%

This is an interim deliverable to demonstrate progress on your project; your script should include comments identifying what each series of steps in the script is intended to do. In general, you’re expected to run some basic descriptive statistics (distributions, means, etc.) that help identify problems with the data or assumptions; if your team has the skills and data for it, you can go a bit further with ANOVA or regression analyses (which will not be covered in class).

As usual, submit the URL to your R script on Canvas.

5.4.5 R Plot Draft 5%

For most projects, some kind of plot is likely to feature heavily in your presentation. Even if not, plotting data is a great skill to develop. It’s also a fairly complex task, so the minimum expectation is that you will use R to create a plot similar to something you might generate in Excel along with an interpretation.

You can submit this deliverable in multiple formats, so choose the one that’s best for your project goals: either create a Markdown file or work ahead and use R Studio to create a knitted HTML document. Most teams will generate at least one plot per research question.

At a minimum, this deliverable must include:

1. One or more plots generated with basic R plotting packages, which must include:
 - Title
 - Labels on axes
 - Readable details like scale values
 - A key or legend, if applicable
2. 1–2 paragraph description of what the plot represents and how it should be interpreted,
3. a link to the commented script for generating the plot/s.

You can earn 1 point of extra credit if you use the `ggplot2` package and “upgrade” your plots (if you use the package but the plots look the same as without, that misses the point).

5.4.6 Audience Persona 5%

You will use the template developed in class to fill in the details for a persona representing your target audience for your analysis. There are two steps to submit this assignment: commit a PDF of the completed draft persona to your project repository and submit the URL for the assignment on Canvas; and reply in the Canvas discussion thread with the URL of the PDF for peer review by Sunday, November 26, at 10 AM.

5.4.7 Presentation 10%

Treat this presentation as a pitch to your clients, a report to your execs, a presentation to city council, or a public speech on the Capitol steps: dress up nice and put on a good show! Additional specifics and requirements related to presentation format will be discussed in class; plan to spend time rehearsing so that your presentation falls within the time limits. Prezi is not permitted.

You will need to address the following points:

- why the questions are important,
- who should care,
- where the data come from,
- how you used it,
- the limitations of the data,
- your interpretation of the data, and
- what decisions should be made based on your analysis.

You should design the presentation to be as persuasive as possible to appeal to your specific audience persona, but without being misleading.

In addition to delivering your presentation in class, submit a URL on Canvas for any materials you use (e.g., slides).

5.4.8 Git Package

15%

Your full Git package documents all the work underlying your project, and will be graded much more closely than other deliverables. You will create a directory in your repository that includes, at a minimum, the following content:

1. Brief but descriptive Readme.md that includes a license for your work, the names of the team members, and an overview of package contents,
2. Fully commented knitted HTML document produced with R Studio that includes:
 - Data cleaning documentation and/or scripts, including data source location,
 - Fully commented analysis script and any outputs that it generates, including summary statistics and plots
3. Audience persona, updated based on peer review feedback, in PDF format
4. A written summary (1000 - 1500 words) in Markdown or PDF format that includes:
 - Your audience and the decisions your analysis targets (which may require some additional background),
 - Brief description of the source data and processing (up to a paragraph each),
 - 2-4 figures (plots) from your analysis,
 - Your interpretations of the plots (up to a paragraph each),
 - A persuasive argument for a decision your specified audience should make based on your results,
 - A word count at the end, not including references, and
 - All relevant references, e.g., data sources and background information.
5. Any presentation slides or materials that you created (PDF format preferred), and
6. A brief Markdown document summarizing contributorship to the individual deliverables for the entire project (who did what), which should be fairly consistent with Git version history.

In other words, it should include all of your other deliverables (the work plan may be omitted) and translate your presentation to a written format, along with a description of how team members contributed to the project. The written summary may be either a Markdown document or PDF. In addition to the knitted HTML file, you will need to include separate files for the Readme, persona, written summary, and contributorship statement, in the formats specified above. The goal is generating a replicable analysis presented as a polished final product that is well organized and easy to understand. Any additional specifications will be discussed in class, and you are encouraged to bring up questions sooner rather than later so unforeseen issues don't trip us all up.

6 Course Readings

Choo, C. W., Detlor, B., & Turnbull, D. (2000). Information seeking on the Web: An integrated model of browsing and searching. *First Monday*, 5(2). <http://journals.uic.edu/ojs/index.php/fm/article/view/729/638>.

- Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2014). *Managing and sharing research data: A guide to good practice* (Chapter 5). Sage.
- Creately.com (2014). *Business Process Modeling Tutorial (BPM Guide Explaining Features)*. <http://creately.com/blog/diagrams/business-process-modeling-tutorial/> Retrieved August 18, 2017.
- Dabbish, L., Stuart, C., Tsay, J., & Herbsleb, J. (2012). Social coding in GitHub: transparency and collaboration in an open software repository. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, 1277–1286. ACM.
- Davis, M. S. (1971). That’s interesting: Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the social sciences*, 1(4), 309–344.
- Georgakopoulos, D., Hornick, M., & Sheth, A. (1995). An overview of workflow management: From process modeling to workflow automation infrastructure. *Distributed and Parallel Databases*, 3(2), 119–153.
- Gershon, N. & Page, W. (2001). What storytelling can do for information visualization. *Communications of the ACM*, 44(8), 31–37.
- Halfaker, A. (2016). Intro to Quarry. https://commons.wikimedia.org/wiki/File:Intro_to_Quarry.pdf
- Hemp, P. (2009, September). Death by information overload. *Harvard Business Review*, 82–89.
- Hullman, J., & Diakopoulos, N. (2011). Visualization rhetoric: Framing effects in narrative visualization. *IEEE transactions on visualization and computer graphics*, 17(12), 2231–2240.
- Lethbridge, T. C., Singer, J., & Forward, A. (2003). How software engineers use documentation: The state of the practice. *IEEE software*, 20(6), 35–39.
- Miaskiewicz, T., & Kozar, K. A. (2011). Personas and user-centered design: How can personas benefit product design processes?. *Design Studies*, 32(5), 417–430.
- Moreau, L., Groth, P., Miles, S., Vazquez-Salceda, J., Ibbotson, J., Jiang, S., ... & Varga, L. (2008). The provenance of electronic data. *Communications of the ACM*, 51(4), 52–58.
- O’Neil, C. (2014). *On Being a Data Skeptic*. O’Reilly Media. Access through Safari Online via the library.
- Quick, J. (n.d.) R Tutorial Series: Summary and Descriptive Statistics. <http://rtutorialseries.blogspot.com/2009/11/r-tutorial-series-summary-and.html>. Retrieved August 22, 2016.
- Thorp, Jer. (2017). *You Say Data, I Say System*. HackerNoon. <https://hackernoon.com/you-say-data-i-say-system-54e84aa7a421>. Retrieved August 16, 2017.
- Tufte, E. R. (1983). *The visual display of quantitative information* (Chapter 2). Cheshire, CT: Graphics press.
- Van den Broeck, J., Cunningham, S. A., Eeckels, R., & Herbst, K. (2005). Data cleaning: detecting, diagnosing, and editing data abnormalities. *PLoS Med*, 2(10), e267. <http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020267>

- Viswanathan, S. & Viswanathan, V. (2015). *R Data Analysis Cookbook*. O'Reilly Media. Access through Safari Online via the Library.
- Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the Belmont Principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 941–953. ACM.
- Welser, H. T., Cosley, D., Kossinets, G., Lin, A., Dokshin, F., Gay, G., & Smith, M. (2011). Finding social roles in Wikipedia. In *Proceedings of the 2011 iConference*, 122–129. ACM.
- Wikimedia Foundation. (2015). <https://meta.wikimedia.org/wiki/Research:Quarry> Retrieved August 22, 2016.
- Wilson, G., Bryan, J., Cranston, K., Kitze, J., Nederbragt, L., & Teal, T. K. (2017). Good enough practices in scientific computing. *PLoS Computational Biology*, 13(6), e1005510.