iSchool at University of Maryland, College Park

INFM 600 **Information Environments** Fall 2016

Course Syllabus as of December 7, 2016

Instructor:	Dr. Andrea Wiggins	E-Mail:	wiggins@umd.edu
Office:	4121G Hornbake	Phone:	301-405-7622
Section 0101:	Thu, 6:00 - 8:45 PM, HBK 0103	Section 0201:	Wed, 6:00 - 8:45 PM, HBK 0109
Class dates:	August 29–December 12, 2016	Office hours:	3:00 - 5:00 PM, Wed & Thu and by appointment

1 Course Description

INFM 600 Information Environments will explore various models and methodologies used to capture and deploy internal and external information and knowledge in a number of settings. Students will analyze organizations in terms of information creation, flow, sharing, conservation, and application to problem solving. The course will take into account both internal and external influences on the management of information and knowledge. We will also examine how information flows and is managed in a variety of settings, and examine a number of examples of successful and unsuccessful online information management.

2 Learning Outcomes

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

1. Find and evaluate data for a purpose
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.
7. Present information for decision support.

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.

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.

Students will need to download and install the following software resources, which are well supported and broadly available:

- GitHub Desktop
- Microsoft Excel (available through TerpWare)
- R Studio

Students will also need to create accounts for the following sites:

- GitHub
- Wikipedia *or* Wikimedia

Assignments will be submitted on Canvas (`http://elms.umd.edu`), but will typically be a URL to a document or repository on GitHub. All online discussions will be hosted on Canvas, and both 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 linked to the weekly module.

3.3 Academic Integrity

The University of Maryland has a nationally recognized Code of Academic Integrity administered by the Student Honor Council. This Code sets standards for academic integrity at Maryland for all undergraduate and graduate students. As a student, you are responsible for upholding these standards for this course. It is very important for you to be aware of the consequences of cheating, fabrication, facilitation, and plagiarism. As defined by the University of Maryland, Academic Dishonesty includes the following activities:

1. "CHEATING: intentionally using or attempting to use unauthorized materials, information, or study aids in any academic exercise.
2. FABRICATION: intentional and unauthorized falsification or invention of any information or citation in an academic exercise.

3. FACILITATING ACADEMIC DISHONESTY: intentionally or knowingly helping or attempting to help another to violate any provision of this Code.

4. PLAGIARISM: intentionally or knowingly representing the words or ideas of another as one?s own in any academic exercise."

Academic dishonesty also includes buying assignments, submitting the same paper more than once, forging signatures, bribery, and other acts that deceive others about your academic work or record. You may also find this Office of Student Conduct definition of academic dishonesty helpful: `http://osc.umd.edu/OSC/AcademicDishonesty.aspx`.

My general policy is "two strikes and you're out." The first incident will be penalized by reduction of up to one letter grade (i.e., 10%). If a second incident occurs, the student will automatically receive a failing grade and will be referred to the Honor Council.

Although these consequences may seem harsh, the consequences for such behavior in a professional setting can be far more devastating to your career and reputation. 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. If you must miss class, notify me in advance by email and check with your classmates afterward so that you can catch up.

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. Since you may need letters of reference for future employment, demonstrating your capacity for professional behavior now is also a great strategy to help ensure that your professors and peers are happy to recommend you for the jobs of your dreams! 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.

3.4.1 Excused Absences

In compliance with University policy, you may excuse yourself from one class session for medical reasons, making a reasonable effort to inform me in advance. More than two absences for medical reasons requires documentation from a health care provider in order to avoid penalties on participation grades.

In addition, it is the student's responsibility to inform me of any intended absences for religious observances within the first two weeks of class (by September 14) to avoid penalties on participation grades.

Students may also be excused for participation in University activities at the request of university authorities; written documentation of such an event is required to avoid penalties on participation grades.

A limited number of make-up credit assignments are available for students who pre-arrange to make up for planned absences. If you know you will have more than one absence, contact me by September 14 to discuss whether this option is available to you.

3.4.2 Inclement Weather

Official closures and delays are announced on the campus website at umd.edu and snow phone line (301-405-SNOW), as well as on local radio and TV stations. Unless there is an official closure or delay, you should assume that class will meet.

3.4.3 Emergency Preparedness

If a public emergency arises, please see the University's Emergency Preparedness Website at `http://www.umd.edu/emergencypreparedness/` for information about the current status of the campus. If a class session needs to be rescheduled, I will email you as soon as possible.

3.5 Communications

Communication outside of class will use Canvas or your umd.edu email account. Course announcements will be posted on Canvas and individual correspondence will be conducted via 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 "INFM600" 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. Telephone is not an effective way to contact me.

3.6 Academic Assistance

If you experience difficulties keeping up with the academic demands of this course, consider contacting the Learning Assistance Service, 2202 Shoemaker Building, 301-314-7693. Their educational counselors can help with time management, reading, math learning skills, note-taking, and exam preparation skills. All services are free to UMD students.

3.7 Disability Accommodations

According to University policy, students with disabilities must acquire documentation from the Disability Support Service Office (4-7682 or dissup@umd.edu) prior to receiving accommodations. However, students are encouraged to speak with me by appointment or during office hours about disability accommodations while awaiting an Accommodation Letter from DSS, which must be presented by the end of the drop/add period.

3.8 Intellectual Property

The University of Maryland's official policy is that copyright for all course materials is held by the professor. Because I hold the intellectual property rights under this policy, my materials are freely available on GitHub under a Creative Commons license.

With respect to ownership of student work, I may request written permission to use exceptional 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.

Please note: Nearly all 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. It 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 14 to discuss alternate options.

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 September 7/8, etc. Please note that some readings are available via URL instead of a PDF, and that required online discussion makes it inadvisable to wait until the last minute to complete course readings.

Week 1, August 31 & September 1: Getting Started

- *Administrivia* Review of syllabus and course policies
- *Topics* Course introduction; Git & and command line basics; Git for collaboration; intellectual property
- *Assignment* Pre-course survey due at start of class
- *Readings* Git install & tutorial; command line basics
- *Guest Lecture* Jonathan Brier

Week 2, September 7 & 8: Information Seeking

- *Topics* Types of data (qual/quant & NOIR); data vs. information vs. knowledge; information seeking & information overload
- *Readings* Ackoff (1999) excerpt; Choo et al. (2000); Hemp (2009)
- *Activities* Information seeking exercises

Week 3, September 14 & 15: Turning Data into Information

- *Topics* Asking questions; where data comes from
- *Readings* Howison et al. (2011); Davis (1971); Evans & Schmalensee (2016)
- *Assignment* Information Seeking - Individual
- *Activities* Developing research questions

Week 4, September 21 & 22: Collaborating with Data

- *Topics* Selecting data for team projects; preventing problems in collaboration
- *Readings* Dabbish et al. (2012); Kotlarsky & Oshri (2005)
- *Activities* Data "pitches" and in-class ranking of team project preferences

Week 5, September 28 & 29: Organizing Information

- *Topics* Metadata, standards, and interoperability; data structures for humans
- *Readings* Corti et al. (2014) Chapter 5; Vardigan et al. (2008)
- *Activities* Data entry analysis
- *Guest Lecture* Jonathan Brier

Week 6, October 5 & 6: Cleaning Data

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

Week 7, October 12 & 13: Relational Databases

- *Topics* SQL basics (SELECT) (+JOIN)
- *Readings* Halfaker (2016); Wikimedia (2015); Stephens & Plew (2002): Hour 1 & Hour 7 (minimum, you may wish to read further); Helland (2016)
- *Assignment* Set up a user account on either Wikipedia or Wikimedia
- *Activities* Querying Quarry; mid-semester course feedback

Week 8, October 19 & 20: Documenting Data Processing

- *Topics* Process documentation in organizations and teams
- *Readings* Lethbridge et al. (2003); Moreau et al. (2008)
- *Assignment* Quarry Queries - Individual; install R Studio
- *Activities* Draft data cleaning process; develop documentation standards

Week 9, October 26 & 27: Basic Descriptive Analysis

- *Topics* R reading in, basic descriptives
- *Readings* Quick (n.d.)
- *Assignment* PBJ Documentation - Individual
- *Activities* R tutorial; PBJ demo; schedule team meetings
- *Guest Lecture* Jonathan Brier

Week 10, November 2 & 3: Project Progress Updates

- *Topics* Team meetings & work session
- *Assignment* Data Cleaning documentation - Team
- *Activities* Team meetings with instructor

Week 11, November 9 & 10: Basic Data Visualization

- *Topics* Basic plotting in R; interpreting descriptive statistics
- *Readings* Hullman & Diakopoulos (2011); Bresciani & Eppler (2008); Tufte (1983) Chapter 2; *optional: Cairo (2015)*
- *Activities* Data interpretation

Week 12, November 16 & 17: Data to Information Workflows

- *Topics* R Studio as a workflow tool; troubleshooting R
- *Readings* Georgakapoulos et al. (1995)
- *Assignment* Draft R Script - Team
- *Activities* R Studio workflow exercise
- *Guest Lecture* Jonathan Brier

November 21—No Class—Thanksgiving

Week 13, November 30 & December 1: Audience Analysis

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

Week 14, December 7 & 8: People's Choice

- *Topics* TBD–make suggestions by email or on mid-semester evaluation
- *Readings* TBD

Week 15, December 14 & 15: Final Presentations

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

5 Assessment

This course provides an overview and introduction to key topics in the field of information management. To practice valuable professional skills, class members will engage in discussions, readings, and collaborative and individual assignments. Discussions will help you develop your ability to reflect about practical issues and discuss these with colleagues. Readings will provide an introduction to topics and exposure to current issues, debates, issues, and solutions. Written and group assignments serve as skill building exercises.

As shown in the table below, 50% of your grade will come from your team projects, 20% 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. Since employers will expect you to speak up and share your insights and expertise, participating in these discussions is good professional practice.

Type	**Assignment**	**Due**	**Points**	**Objectives**
Individual	Pre-course Survey	Week 1	N/A	
Individual	Info Seeking	Week 3	10	1, 2, 3, 7
Individual	Quarry Queries	Week 8	10	3
Individual	PBJ Documentation	Week 9	10	5
Team	Work Plan	Week 6	5	2
Team	Data Cleaning	Week 10	5	4,5
Team	Progress Meeting	Week 10	5	2
Team	*Draft* R Script	Week 12	5	5, 6
Team	*Draft* R Plot	Week 13	5	5, 6
Team	Project Presentation	Week 15	10	7
Team	Git Repository	Finals	15	all
Participation	In-class Participation	throughout	15	all
Participation	Online Discussion	throughout	5	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. Failure to meet document requirements will be penalized as specified in the rubrics for assignments; content that does not match formatting requirements will be subject to additional scrutiny for potential plagiarism.

Assignments that do not follow these specific requirements (where applicable) will lose points on the assignment grade. In addition to punctuality, the grammar, presentation and your ability to follow instructions are very important, as in the real world, so it is essential that you spell check and proofread your documents. For written assignments, proofreading a *printed* copy of your work

is especially effective for finding errors that you might overlook on the screen. Note that standard Word document templates do not meet the criteria below.

Document requirements:

- Use **11pt Times New Roman** for body text.
- All text must be in **black**, without highlighting or background colors. Be very careful what you copy and paste!
- Documents must use **1.5 line spacing with 1" margins on all sides in 8.5" x 11" (US letter) format**.
- All written content must have a blank line between paragraphs (block format) OR the first line of the paragraph must be indented.
- You may use larger font sizes, sans serif fonts, boldface, and/or italics for title text and section or table headers, but it must still be in black type.
- On every page, document headers must include your **name** and **UMD email address or team name** on the *left* and **page numbering** on the *right*. Exception: for the written portion of the Team Project, cover pages should not include numbers or headers.
- At the end of the document, insert the **word count**, not including references, appendices, or executive summary (where applicable).
- 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.
- Most assignments can be submitted in multiple formats; see assignment details for specifics.

These requirements do not apply to R outputs, Markdown documents, or Git repositories. You may choose to submit a Markdown document on Git instead of a word processing document, but it must be nicely formatted and easy to read. The final commit on Git documents must be timestamped before 6 PM on the date the assignment is due; the last commit prior to the cutoff time will be graded.

When you prepare assignments or post on the discussion boards be sure to provide proper bibliographical information for any sources referenced, for direct quotations, and for the sources of key concepts or ideas. Check the UMD citation guide for more details: `http://www.lib.umd.edu/ues/guides/citation-tools`.

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 so 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.

100%+: A+
96 - 99%: A (4.0)
92 - 95%: A- (3.7)
88 - 91%: B+ (3.3)
84 - 87%: B (3.0)
80 - 83%: 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 Discussion and Participation — 20% total

5.3.1 In-class Participation — 15%

Class discussions are an important way to learn and demonstrate learning; 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, however, and you are expected to actively engage in course activities and discussions. You can earn 1 point (percent) each week, with an additional point for actively interacting with the presentations on the last day.

5.3.2 Online Discussions — 5%

Throughout the semester, we will have a few online discussions to extend our classroom conversations and reflect on complex issues. At least 6 discussion forums will be posted; you must respond to at least 5 different discussions. If you submit a post for more than 5 online discussions *and* all posts make an adequate contribution, you can earn **up to 1 point of extra credit**. An "adequate contribution" is a nontrivial post that conveys more than a simple agree or disagree: your goal is to make an interesting point, bring up a worthwhile question, point to an informative related resource, and constructively contribute to the conversation. Most posts can accomplish this goal in about a paragraph, which should be clearly and professionally written. If you struggle with spelling or grammar, compose your post in a word processor to use the grammar and spelling tools before pasting the text into Canvas. Discussions will be open for one week.

5.4 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.

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:

- 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 exercises; 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.
- Data that contain personally-identifiable information such as full names, email addresses, contact details, etc.
- 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, Twitter, or Facebook.

Note that some of the most interesting data may be available to use by permission but not fully available online (e.g., UMD's Office of Sustainability has very interesting data, only some of which is online.) 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, provide the following information in the following order:

1. An APA-formatted data citation (including URL)
2. Details of the license or terms of use (include a link if needed)
3. About one paragraph describing why these data are interesting
4. Potential data users and decision-makers for this data
5. Three questions this data might help to answer; note additional sources needed if applicable

Compose a word-processing document and commit to Git, or create a Git markdown document with the required details; submit the URL to the document in the Git repository on Canvas.

A few students in each class 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 to introduce the class to the data for consideration for team projects.

5.5 Quarry Queries — 10% total

This assignment provides practice with SQL and interpreting data, as well as exposure to how peoples' work is represented in data and provenance. The further back you browse in the Quarry query history, the more likely you are to find some good examples. You can work together on this assignment to make a little more sense of it, so long as you each pick your own sets of queries to

examine and discuss. There is more than enough Quarry history for every student to work with unique sets of queries. *Any assignments with identical query sets will be carefully scrutinized for academic dishonesty.*

- Find a series of 2+ queries by a single user and try to figure out what they were working on. You may have to be a little imaginative with this; make your best guess based on what you see. A few additional tips:
  - You can also join the IRC channel to ask for help translating details at `http://webchat.freenode.net` with channel #wikimedia-research. Especially helpful users include: yuvipanda, halfak, & J-Mo.
  - Any version of "wiki" that doesn't start with "en" represents a non-English Wikipedia (e.g., hewiki is Hebrew Wikipedia).
  - If the query does NOT start with "use XXwiki..." then it defaults to enwiki, the English Wikipedia.
  - Usually sequential queries represent either refining or drilling down into a single task with multiple approaches, or carrying out a set of tasks that may or may not be tightly related (but might, for instance, all deal with maintenance on Hebrew Wikipedia).
  - You can also try googling some pieces of the queries to find documentation, and check the "Discuss" tab to see if anything there relates to what you're looking at.
- Paste in the URLs from the series of queries you found and write up a short description (2–3 sentences) of a plausible explanation for the work that the Quarry user might have been doing.
  - Some of the query sets are more obvious or opaque than others, so don't hesitate to shop around for series of queries before settling on an example.
  - Another option for finding good queries is to browse user profiles. Some of the users that appear frequently or have nicely titled queries may have better examples for you to check out.
- Repeat this process for 3 sets of queries; each set of queries can come from different users.

On Canvas, submit the URL for a word processing or markdown document on Git that contains the details above for the three sets of queries that you identified.

5.6 PBJ Documentation — 10% total

For this assignment, you will develop detailed, step-by-step process documentation of how to make a quintessentially American delicacy: a peanut butter and jelly (or jam) sandwich. Your process documentation must be explicit and comprehensive, so leave no room for confusion. While this task might initially appear simple, it offers a good example of the challenges of technical writing; if you have questions, ask! It will also be a lot more fun and interesting if you do *not* search the web for examples. As always, make sure to proofread your documentation for spelling and grammar to receive full credit.

On Canvas, submit the URL for a word processing or Markdown document on Git.

5.7 Team Project — 50% 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 that there is plenty to learn from the experience of developing a fully documented start-to-finish data-driven decision support deliverable. In the process, you will also gain familiarity with 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 on details that might otherwise derail you. 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 the work plan, data cleaning document, team progress meeting, R script draft, and R plot draft. That's 25% of your grade!

Here's the hardest part: the sophistication of your analysis is up to you! While we know that the level of uncertainty adds to the challenge, we also know you're capable of handling it, and you'll hear about it if your plans are off-target. Challenge yourself just enough to learn something new and fulfill the spirit of the assignment, but don't let it drag you under. If your planned analyses are too trivial or too complex, we will discuss more suitable options during the Progress Meeting in Week 10, which should leave enough time to level up or down as needed. If you need feedback before Week 10, please drop in during office hours or make an appointment.

This project will include:

1. Work plan: Week 6, 5 points
2. Data cleaning documentation draft: Week 10, 5 points
3. Team progress meeting: Week 10, 5 points
4. R script draft: Week 12, 5 points
5. R plot draft: Week 13, 5 points
6. Presentation: Week 15, 10 points
7. Git package: Week 15, 15 points

We will take advantage of your work on the Information Seeking assignment to identify viable data sets and assign them to groups based on preferences. Groups will also be assigned by the instructors to ensure a fair distribution of pre-existing skill sets. To make these processes work smoothly, please be sure to respond to the pre-course survey promptly!

5.7.1 Work Plan & Team Progress Meeting — 5% & 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 team progress meeting with the instructor, which 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 work plan should be a simple document (word processing or Markdown) containing the items mentioned above and made available from your team's Git repository. Submit the URL for the work plan on Canvas.

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

5.7.2 Data Cleaning & Documentation 5%

Your data cleaning and documentation draft will be part of your final project repository; this is due by Week 10 to help provide early feedback. 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.
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 us help you troubleshoot. Any issues highlighted by instructor feedback should be carefully attended to for your final project data processing documentation. This document can take several forms (R script, Markdown, word processing), so choose the best one for your project needs and submit the URL on Canvas.

5.7.3 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). Again, this is a draft deliverable, so it's a good opportunity for feedback if you're getting stuck or need suggestions on scaling up or scaling back your planned analysis. In particular, if there are issues or questions you'd like to see addressed in class, we will have a Canvas discussion board open so

you can nominate and vote on topics to cover in Week 12. Chances are, if you're running into a problem, others are too, so don't be afraid to share your questions or challenges—you're doing us all a favor!

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

5.7.4 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 start developing! 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.

Again, you can submit this deliverable in multiple formats; choose the one that's best for your project goals. You can paste a plot and description into a word processing or Markdown file, for example. At a minimum, this deliverable must include:

1. A plot 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.

You can earn 1 point of extra credit if you use the ggplot2 package; *however*, if ggplot2 is too complicated for your team, don't do it—it's not worth what you'd lose for a late or incomplete deliverable! You can earn 1.5 points of extra credit if you use the xkcd package on top of ggplot2, but the same warning (and then some!) applies.

5.7.5 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! You will have 8 minutes to present your work, with 2 minutes for questions. Rehearse in advance, as the timer will go off at 8 minutes and won't stop until you do. Additional specifics and requirements related to presentation format will be discussed in class.

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,
- 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 without being misleading.

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

5.7.6 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. Fully commented data cleaning documentation and/or scripts, including data source location,
2. Fully commented analysis script and any outputs that it generates, which can be combined with the data cleaning document,
3. A written summary (up to 500 words) that includes:
   - Your audience and the decisions your analysis targets,
   - Brief description of the source data and processing (up to a paragraph each),
   - A figure (plot) from your analysis,
   - Your interpretation of the plot (up to a paragraph),
   - A persuasive argument for a decision your audience should make based on your results,
4. Any presentation slides or materials that you created, and
5. A brief document summarizing contributorship to the individual project 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 (except the work plan) and translate your presentation to a written format, along with a description of how team members contributed to the project. This can all be accomplished with a single document generated by R Studio (preferred); however, you can also create a set of files. 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

Bresciani, S., & Eppler, M. J. (2009). The risks of visualization: A Classification of the Disadvantages Associated with Graphic Representations of Data. *Identitat und Vielfalt der Kommunikations-wissenschaft*, 165–178.

Cairo, A. (2015). Graphics lies, misleading visuals. In *New Challenges for Data Design* (Chapter 5), 103–116. Springer London.

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.

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.

Evans, D. S. & Schmalensee, R. (2016). How We Learned (Almost) Everything That's Wrong with U.S. Census Data. *Harvard Business Review*.

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`

Helland, P. (2016). The Singular Success of SQL. *Communications of the ACM*, 59(8), 38–41. DOI: 10.1145/2948983.

Hemp, P. (2009, September). Death by information overload. *Harvard Business Review*, 82–89.

Howison, J., Wiggins, A., & Crowston, K. (2011). Validity issues in the use of social network analysis with digital trace data. *Journal of the Association for Information Systems*, 12(12), 767–797.

Hullman, J., & Diakopoulos, N. (2011). Visualization rhetoric: Framing effects in narrative visualization. *IEEE transactions on visualization and computer graphics*, 17(12), 2231–2240.

Kotlarsky, J., & Oshri, I. (2005). Social ties, knowledge sharing and successful collaboration in globally distributed system development projects. *European Journal of Information Systems*, 14(1), 37–48.

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.

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.

Stephens, R. & Plew, R. (2002). *Sams Teach Yourself SQL in 24 Hours, 3rd Edition*. Pearson. `http://www.informit.com/library/library.aspx?b=STY_Sql_24hours`. Hour 1 & Hour 7 (minimum). Retrieved August 22, 2016.

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`

Vardigan, M., Heus, P., & Thomas, W. (2008). Data documentation initiative: Toward a standard for the social sciences. *International Journal of Digital Curation*, 3(1), 107-113.

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.

Wikimedia Foundation. (2015). `https://meta.wikimedia.org/wiki/Research:Quarry` Retrieved August 22, 2016.