Data Science and Visualization for Policy Analysis
INTA 708

Fall 2021
Wednesdays 1330 - 1620
1055 Allen Building

Instructor: Jesse H. Sowell II

Office: 1096 Allen Building

E-mail: jsowell@tamu.edu

Office Hours: Tuesdays 1630 - 1800 by appointment

Course Description

For many policy issues—in particular those at the intersection of science, technology, and public policy—the analyst is faced with a surfeit of data and the challenge of transforming this data into a form that can be tractibly evaluated. This course equips students with the data science skills necessary to make sense of our increasingly data rich policy environment. Starting with exploratory analysis driven by data visualization, students will develop the skills and workflows necessary to identify and develop compelling visualizations of trends in domestic and transnational policy issues.

As per our primary text (Wickham, 2017), data analysis falls into two general groups: hypothesis *generation* and hypothesis *confirmation*. This course focuses on data science and visualization for hypothesis *generation*. Part I focuses on what is colloquially referred to as "data wrangling": importing, cleaning, and integrating data from a variety of sources. Part II introduces the fundamentals of exploratory data analysis (EDA), visualization tools, mapping (basic GIS), and clustering, concluding with text mining. Following the theme of this course, we focus on how to use, interpret, and visualize analyses generated using these tools and strategies.

Course Schedule Overview

Week	Date	Lecture
01	01 September 2021	Introduction to Data Science for Policy Analysis
02	08 September 2021	Introduction to R
03	15 September 2021	Cleaning, Transforming, Managing Your Data
04	22 September 2021	Grammar of Graphics I
05	29 September 2021	Grammar of Graphics II
06	06 October 2021	Exploratory Data Analysis
07	13 October 2021	Reading Week
08	20 October 2021	Exploratory Data Analysis II
09	27 October 2021	Mapping
10	03 November 2021	Cluster Analysis I
11	10 November 2021	Cluster Analysis II
12	17 November 2021	Text Mining I
13	24 November 2021	Thanksgiving
14	01 December 2021	Text Mining II
15	08 December 2021	Project Presentations

Course Prerequisites

There are no prerequisites for this course. A background in basic statistics is strongly recommended.

Course Learning Outcomes

- understand the philosophy of data science applied to exploratory data analysis (EDA) and hypothesis *generation*
- apply this philosophy to a mixed methods approach to data science and visualization
- master data management skills, in particular the extract, load, transform (ELT) process
- understand the core principles of "data wrangling" using the principles of data transforms and the attendant tools in the `tidyverse` family of packages
- be able to evaluate visualizations in the media in terms of mapping data to aesthetics, as presented by Healy
- understand the principles of the grammar of graphics and apply these to develop visualizations
- understand and develop diagnostic, analytic, and public facing visualizations
- apply data transformations and visualization tools to perform common EDAs such as evaluating common aggregate indicators (measures of centrality and distributions) and develop dataset specific EDAs
- develop basic geographic information system (GIS) and mapping skills that build on the application of data aesthetics and the grammar of graphics to evaluate and visualize spatial trends
- understand the principles behind common clustering algorithms (k-means and hierarchical), how these facilitate identifying trends, patterns, and groupings of observations
- apply clustering to both quantitative and categorical variables
- understand how to apply, compare, and evaluate the results of different clustering strategies, and the kinds of groups that can be identified by those strategies
- apply principle component analysis (PCA) to reduce the dimensionality of clusters and effectively visually inspect, compare, and analyze groups identified by various cluster analyses
- use text mining and the principles of clustering to evaluate groups of text documents (corpi) for trends in concepts and sentiment
- use frequency and network analysis to better understand the relationships between concepts within and across corpi

Textbooks and Resource Materials

Books

Required

Each of these texts are freely available online or will be provided as PDF in the Zotero course library.

Grolemund, Garrett (2014). *Hands-On Programming with R: Write Your Own Functions and Simulations*. 1 edition. https://learning.oreilly.com/library/view/hands-on-programming-with/9781449359089/. O'Reilly Media.

Healy, Kieran (2018). *Data Visualization: A Practical Introduction*. 1st edition. https://socviz.co/. Princeton, NJ: Princeton University Press.

Lindgren, Simon (2020). *Data Theory: Interpretive Sociology and Computational Methods*. https://www.amazon.com/Data-Theory-Interpretive-Sociology-Computational-ebook-dp-B08HCGBH9L/dp/B08HCGBH9L/. Wiley.

Silge, Julia and David Robinson (2017). *Text Mining with R: A Tidy Approach*. 1st edition. https://www.tidytextmining.com/. O'Reilly Media.

Wickham, Hadley and Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1st edition. https://r4ds.had.co.nz/. O'Reilly Media.

References

The following are optional reference materials, some of which will be used as source materials for lectures.

Bruce, Peter, Andrew Bruce, and Peter Gedeck (2020). *Practical Statistics for Data Scientists, 2nd Edition.* Second. https://learning.oreilly.com/library/view/practical-statistics-for/9781492072935/. O'Reilly Media. *(electronic version available from library)*

Dalgaard, Peter (2002). *Introductory Statistics with R.* New York, NY: Springer. *(electronic version available from library)*

Engel, Claudia A. (2019). *Using Spatial Data with R.* https://cengel.github.io/R-spatial/. *(in Zotero, available online)*

Imai, Kosuke (2018). *Quantitative Social Science: An Introduction.* Princeton: Princeton University Press. *(PSEL library reserve, Dr. Sowell also has a copy)*

(in Zotero, available online)

Peng, Roger D. (2016). *Exploratory Data Analysis with R.* https://bookdown.org/rdpeng/exdata/. *(in Zotero, available online)*

(in Zotero)

Xie, Yihui, J. J. Allaire, and Garrett Grolemund (2019). *R Markdown: The Definitive Guide.* https://bookdown.org/yihui/rmarkdown/. CRC Press. *(in Zotero, available online)*

Xie, Yihui, Christophe Dervieux, and Emily Riederer (2021). *R Markdown Cookbook.* https://bookdown.org/yihui/rmarkdown-cookbook/. *(in Zotero, available online)*

Course Tools

The following course tools should be installed **before** the first lecture on Wednesday 01 September 2021.

R and RStudio

We will be using R and RStudio for data analysis and visualization. You will need to download and install the following applications to complete the assignments in this course.

1. R at https://cran.revolutionanalytics.com/
2. RStudio at https://www.rstudio.com/products/rstudio/download/#download

Students should install these applications in the order above. Please be sure to install these tools before the first class. We will be using RStudio extensively in each class to demonstrate various tools, methods, and visualization techniques. We will introduce some of the foundations of using RStudio in the first lecture, introducing additional skills and tools as the course progresses.

Zotero

We will be using Zotero to access course materials and to manage the references used in the course assignments. Unless otherwise indicated, students can find any of the materials listed on this syllabus (with the exception of some of the textbooks above) in the shared Zotero library for this course. These materials include journal articles, conference papers, newspaper and magazine articles, **lecture slides**, and **the most up-to-date version of this syllabus**. Dr. Sowell will be sending invitations to the Zotero shared library after the first lecture, the afternoon of Wednesday 01 September 2021. If the student has not received an invitation to the shared library, check your spam folder. If the student still cannot find the invitation, e-mail Dr. Sowell (jsowell@tamu.edu).

The first step to using Zotero is to create a Zotero account. Students can download the Zotero app at https://www.zotero.org/download/. Students should also install the Zotero Connector for the browser of their choice. For step-by-step instructions, see the section on Zotero Configuration in the Appendix. Word

processor plugins are available for Word, LibreOffice, and Google Docs. That said, in this course we will be doing all the write-ups in Rmarkdown, so you will not need these.

TAMU libraries offers extensive documentation and tutorials on using Zotero. Please see:

- TAMU Zotero Research Guide
- Creating Bibliographies, in particular, the *two-minute* video that shows how to insert in-text citations into a Word document and how to generate bibliographies.
- The *less than two-minute* quick guide video for saving citations from your web browser

It should take less than 30 minutes to get the Zotero app and connector installed, setup, and then run through the two video guides. This will save you many more hours fiddling with references when writing your policy research projects.

Finally, to incorporate references into your mini-project and policy research project submissions, you will need to install the Zotero plugin Better BibTeX. It looks very technical, but you can ignore that for now. You can find the link for installing Better BibTeX at https://retorque.re/zotero-better-bibtex/installation/. We will work through the nuance of configuration in lecture.

Lecture slides will be added to the shared library at latest one hour before each class. The syllabus and class readings will be periodically updated with contemporary readings from the news related to upcoming topics in the course. To ensure you have the latest syllabus, it is strongly suggested that you open the syllabus directly from Zotero.

To be clear on the locations of these materials:

- the latest syllabus can always be found in the directory `INTA 689 - Data Science/Syllabus` (Zotero calls directories *collections*) of the shared library
- slides will be in the collection `INTA 689 - Data Science/Classes/ClassX/Slides` where `X` is the class number (01, 08, 12, etc.)
- references in slides that are not from one of the assigned books or one of the readings lists, can also be found in the `Slides` collection for that lecture

Please contact Dr. Sowell (jsowell@tamu.edu) if you have any problems accessing Zotero or the class materials in the shared library INTA 689 - Data Science.

GitHub

We will be using Github for your projects and assignments. In terms of software, students should

1. create a GitHub account using their `@tamu.edu` e-mail address
2. download GitHub Desktop, install it, and log into their GitHub account with the app

In the first class we will do a brief tutorial on how students will use GitHub to manage their policy research projects and learning exercises.

**All project milestones and learning exercises for this course will be submitted via GitHub.**

Grading Policy and Overview of Assignments

This is a project-based course, there are no exams. Details and guidelines for each assignment are provided in the Assignments section. Final grades will be calculated as follows:

1. participation, 10%
2. learning exercises, 30%
    - LE-01, Learning R, *due Monday 20 September 2021*
    - LE-02, Transforms, *due Monday 11 October 2021*
    - LE-03, Visualization and EDA, *due Monday 08 November 2021*
    - LE-04, Clustering, *due Monday 22 November 2021*
    - LE-05, Mapping, *due Monday 06 December 2021*

3. policy research project, total 60%, breakdown:
   - setup shared policy research project library in Zotero, 3 *participation points*, *due Friday 03 September 2021*
   - proposal, 5%, *due Monday 20 September 2021*
   - detailed outline and exploratory data analysis, 20%, *due Monday 15 November 2021*
   - in class presentation, 5%, *due Tuesday 07 December 2021*
   - policy research project report, 30%, *due Wednesday 08 December 2021*

Class Milestones and Exercises Timeline
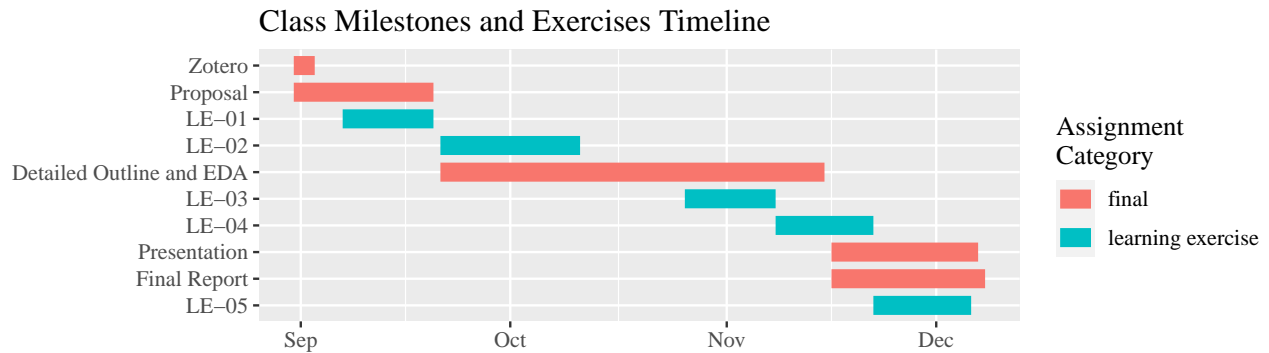
Grades for assignments will be in terms of total points for the class. For instance, a good grade for the proposal would be 4.75/5.

Final letter grades will be assigned as follows:

letter grade	range
A	>= 90%
B	>= 80%, < 90%
C	>= 70%, < 80%
F	< 70%

In terms of evaluation, grades for project milestones (within the scope of the assignment) are assessed as follows:

- **A+, >= 96%** indicates
  - exceptional mastery of concepts at hand,
  - exceptional application of the concepts,
  - salient issues and concepts covered in the class are addressed,
  - appropriate trade-offs are discussed,
  - analysis is supplemented by contemporary instances of the problem from outside materials,
  - exceptional articulation, with an introduction to the problem, challenges, trade-offs, and recommendations where requested
- **A, >= 90%, <96%** indicates
  - accurate articulation of concepts at hand,
  - effective applicatin of the concepts,
  - *most* salient issues and concepts covered in the class are addressed,
  - appropriate trade-offs are discussed,
  - good articulation of the analysis with an introduction to the problem, challenges, trade-offs, and recommendations where requested
- **B, >= 80%, < 90%** indicates
  - accurate articulation of the concepts at hand,
  - effective application of the concepts,
  - only *some* key issues and concepts related the problem at hand are presented,
  - some trade-offs discussed in class are missing,
  - weak articulation of analysis, has only rudimentary introduction to the problem, challenges, trade-offs, and recommendations where requested
- **C, >= 70%, < 80%** indicates

– inaccurate articulation of the concepts at hand,
– weak or unclear application of the concepts,
– significant key issues and concepts related to the problem at hand are missing or misconstrued,
– limited discussion of trade-offs,
– poor articulation of analysis, does not have a clear introduction to the problem, challenges, trade-offs, and recommendations where requested
– **F, < 70%** indicates
– inaccurate representation of the concepts at hand,
– little to no application of the concepts,
– signifiant number of the key issues and concepts related to the problem at hand are missing or misconstrued,
– very little discussion of trade-offs or single-sided,
– writing is unclear and unstructured

Specific criteria for individual milestones can be found in the assignment descriptions.

Each set of learning exercises has an assigned point value. Each problem within the learning exercises has a clearly indicated point value. Learning exercises are evaluated in terms of the completeness and correctness of the solution to each problem. If no solution is offered, that problem is assessed at zero points. Partial solutions will garner points for that problem based on how complete and correct the solution is. Correct solutions (not always exactly the same as the solution) will garner all the points for that problem.

Late Work Policy

Enforcement of the following late work policy is at the discretion of the instructor.

As noted in the discussion of GitHub, *all assignments for this class are due at or before 2359 on the due date for the assignment.* **Late submissions will incur a penalty of 10% per day after the due date.** For instance, if the assignment is due on Monday and it is submitted on Wednesday, a 20% late penalty will be applied. *Assignments submitted on or after the tenth day after the deadline will receive zero points and will not be graded.*

Ideally, everyone plans ahead and gets work done ahead of time. That said, every student gets one *late submission mulligan.* Everyone gets behind at some point or finds they have several deliverables due at the same time and needs a little slack. If you see yourself heading for this situation, to use your mulligan you must e-mail the instructor *at least 24 hours before the deadline* to indicate you would like to take your mulligan, explain why, and when you think you can submit the work. The instructor will work with you to identify a reasonable revised deadline. You will likely get a day or two more time, a week is unacceptable with the exception of dire circumstances. *Like the overall late policy, the mulligan is also at the discretion of the instructor, so please do not abuse this generous option.*

Course Schedule: Lectures and Readings

All of the assigned textbooks, journal articles, papers, newspaper articles, and other documents assigned in the reading lists below can be found in the shared Zotero library for this course. Lecture slides will be added to the appropriate Zotero shared folder at latest one hour before class.

For any given class there will be at most three readings lists:

– **Essential Readings** are the *required* readings from the textbooks and course materials.
– **Contemporary Readings** are *strongly recommended* after reading the essential readings.
– **Optional Readings** are *not required.* These readings may be referenced in lectures. Optional readings may also be useful starting points for policy research project research.

Class #01: Wednesday 01 September 2021

Introduction to Data Science for Policy Analysis

In this lecture we will discuss the course objectives, get a feel for where each student is in terms of experience with data analysis and statistics, and review the course timeline and objectives. We will briefly review the difference between hypothesis generation versus hypothesis validation—this class largely about data science and visualization in support of hypothesis *generation*, although those with a sufficient background may incorporate validation into their projects. In the second half of the class we will introduce the tools we will be using in the class: the concepts and principles behind analysis in R, RStudio as the primary analysis tool, and RMarkdown as the tool for literate, reproducible research.

Essential Readings

- **Numanović, Amar (2017).** ***Data Science: The Next Frontier for Data-Driven Policy Making?*** https://medium.com/@numanovicamar/https-medium-com-numanovicamar-data-science-the-next-frontier-for-data-driven-policy-making-8abe98159748.

- **Wang, Joan (2017).** ***Musings at the Intersection of Data Science and Public Policy.*** https://towardsdatascience.com/musings-at-the-intersection-of-data-science-and-public-policy-cf0bb2fadc01.

Part I: Data Wrangling

Class #02: Wednesday 08 September 2021

Introduction to R

Data Science is about exploration and prediction. In this first half of this class, we will focus on exploration, in particular, methods and strategies for hypothesis *generation*. In this class we will provide an overview of the difference and relationship between hypothesis generation and confirmation and how these different strategies contribute to policy analysis. We will illustrate the concepts with instances we will explore in depth in class. In this class we will also introduce the basics of R: types of variables, how data is stored in R. We will illustrate these concepts through numerous examples from the text and the class datasets.

Essential Readings

- **Gastner, Michael T. (2021).** ***Data Analysis and Visualisation with R.*** http://michaelgastner.com/DAVisR2021/.

  Read Chapters 2; 3; Chapter 8, Sections 8.1(http://michaelgastner.com/DAVisR2021/chap-packages.html#installing-a-package) and 8.2; Section Section 4.1; Chapters 11 and 12.

  Gastner provides an excellent introduction to RStudio (Chapter 2), along with an introduction to the basic data types such as vectors (Chapter 3) and factors (Chapter 12). Factors are always one of the most challenging of the basic types for students to get used to, so you should spend some time to make sure you understand how they work.

- **Wickham, Hadley and Garrett Grolemund (2017).** ***R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.*** **1st edition.** https://r4ds.had.co.nz/. **O'Reilly Media.**

  Read Chapters 1, 2, and 4.

  These Chapters introduce you to some of the language used in Wickham and some of the principles of data science. We will return to the data science workflows introduced in these chapters throughout the course.

- **Healy, Kieran (2018).** ***Data Visualization: A Practical Introduction.*** **1st edition.** https://socviz.co/. **Princeton, NJ: Princeton University Press.**

Read Chapter 2, Appendix 1.

Pay special attention to Section 2.4, be patient and realize the only way to become effective with these tools is practice.

– **Wickham, Hadley and Garrett Grolemund (2017).** *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* **1st edition.** https://r4ds.had.co.nz/. **O'Reilly Media.**

Read Chapter 27.

This chapter introduces you to RMarkdown. As Wickham highlights, the best way to learn is to play with a few documents and try things out. All of your learning exercises and your policy research project are in RMarkdown, so you will have plenty of additional opportunities to apply these skills. This chapter should also serve as your reference to RMarkdown.

– **Cone, Matt (2019).** *Markdown Guide: Basic Syntax.* https://www.markdownguide.org/basic-syntax/.

This is the foundational syntax for RMarkdown, the tool we will be using to create easily and systematically reproducible notebooks and reports. We will walk through some simple examples to get you started in the second half of lecture.

Optional Readings

– **Gastner, Michael T. (2021).** *Data Analysis and Visualisation with R.* http://michaelgastner.com/DAVisR2021/.

Read Chapter 15.

We will go over how to read data from `.csv`, `.xls[x]` files, and from Google Sheets in the class slides, but if you would like a reference, see Gastner, Chapter 15. This is assigned reading for the next class, but included here if you want to jump ahead a smidge.

– **Ramirez, Rebecca (2020).** *The Science Behind Storytelling.* https://www.npr.org/2020/08/18/903545336/the-science-behind-storytelling.

This is a fun podcast on the role of storytelling in science communication, appropriate for how we are framing the hypothesis generation process as a form of storytelling with data and visualization.

– **Xie, Yihui, J. J. Allaire, and Garrett Grolemund (2019).** *R Markdown: The Definitive Guide.* https://bookdown.org/yihui/rmarkdown/. **CRC Press.**

– **Xie, Yihui (2019).** *Bookdown: Authoring Books and Technical Documents with R Markdown.* https://bookdown.org/yihui/bookdown/. **CRC Press.**

These are good references when you want to dig deeper into the nuanced options and formats available in RMarkdown.

Class #03: Wednesday 15 September 2021

Cleaning, Transforming, Managing Your Data

One of the most difficult and potenially daunting element of data science is data management. In this class we will introduce data cleaning and transformation tools and strategies for making this task easier. Data cleaning is fundamentally about consistency: we will look at how to systematically identify and correct for inconsistencies and missing data. Data transformation is where we start to manipulate the data to identify trends by filtering to focus on select subsets of your data, sorting, selecting particular variables to focus on, creating new variables from existing variables, and grouping your data. The class concludes with a discussion of how combinations of these basic transforms contribute to hypothesis generation.

Essential Readings

– **Healy, Kieran (2018).** *Data Visualization: A Practical Introduction.* **1st edition.** https://socviz.co/. **Princeton, NJ: Princeton University Press.**

Read Chapter 1.

Why are we reading about visualization in the transforms lecture? We want to start thinking about the implications of transforms for how we are going to visualize the data later. It is an iterative process, but having an idea of where one is going saves some (but not all) trial and error.

– **Wickham, Hadley and Garrett Grolemund (2017).** *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* **1st edition.** https://r4ds.had.co.nz/. **O'Reilly Media.**

Read Chapter 10.

The main rectangular data structure we will be working with most of the semester is called a `tibble`. It is what is more broadly referred to as a `data frame`. As you will see in the readings, when displayed it much looks like a spreadsheet. There is a lot of nuance regarding how you can create `tibble`s and in this Chapter Wickham provides some of the most common.

– **Gastner, Michael T. (2021).** *Data Analysis and Visualisation with R.* http://michaelgastner.com/DAVisR2021/.

Read Chapter 14 and 15.

Chapter 14 of Gastner further illustrates illustrates the construction of `tibble`s. Chapter 15 illustrates how to read data into a tibble from `.csv` and `.xls[x]` files.

– **Wickham, Hadley and Garrett Grolemund (2017).** *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* **1st edition.** https://r4ds.had.co.nz/. **O'Reilly Media.**

Read Chapter 5.

Part II: Exploratory Data Analysis

Class #04: Wednesday 22 September 2021

Grammar of Graphics I

Data visualization is part art, part science. This class will introduce you to the grammar of graphics, a modular system of constructing data visualizations. For this class, the objective is to develop an intuition for how different aesthetics are used to highlight trends in your data. It should be stressed this class is intended to develop an intuition of how the grammar of graphics works, we be refining these skills throughout the course.

Essential Readings

– **Healy, Kieran (2018).** *Data Visualization: A Practical Introduction.* **1st edition.** https://socviz.co/. **Princeton, NJ: Princeton University Press.**

Read Chapter 3.

You will notice that Healy does not use our friend the pipe function `%>%` as extensively as we do in this chapter. As we learned in the last lecture, the pipe is our friend. That said, Healy illustrates how to make modular, composable visualizations, the foundation of dynamic, automated generation of families of analyses and visualizations. You will develop an intuition about when to modularize and when to use large compositions as illustrated by Wickham as you gain more experience.

– **Wickham, Hadley and Garrett Grolemund (2017).** *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* **1st edition.** https://r4ds.had.co.nz/. **O'Reilly Media.**

*Read Chapter 3, **pay close attention** to the concepts of aesthetics and geoms.*

Optional Readings

– **Wickham, Hadley (2010). "A Layered Grammar of Graphics''. In:** *Journal of Computational and Graphical Statistics* **19.1, pp. 3–28. DOI: 10.1198/jcgs.2009.07098.**

Many of the concepts laid out in the essential reading are elaborated in technical depth here. It is worth at least a skim

– **Wickham, Hadley (2019).** *Ggplot2: Elegant Graphics for Data Analysis.* **Third.** https://ggplot2-book.org/. **Springer.**

The authoritative reference for `ggplot2` and applying the grammar of graphics.

Class #05: Wednesday 29 September 2021

Grammar of Graphics II

In this lecture we will dig a bit more into the grammar of graphics. Following Healy Chapter 4, we will dive into a few more examples of how aesthetic mappings work, further developing our intuition around the grammar of graphics. Chapter 5 integrates what we have learned about transforms in Lecture 3 with how `ggplot` and the grammar of graphics transform data. In this second half we will focus on how to prepare your analyses for visualization. Finally, we will introduce additional `geom`s, tools for mixing these, and how to better control information about your visualizations in terms of guides and legends.

Essential Readings

– **Healy, Kieran (2018).** *Data Visualization: A Practical Introduction.* **1st edition.** https://socviz.co/. **Princeton, NJ: Princeton University Press.**

Read Chapters 4 and 5.

Class #06: Wednesday 06 October 2021

Exploratory Data Analysis

In this class we will delve into the art and science of Exploratory Data Analysis (EDA). In this first half of the module will highlight EDA as an iterative, creative process of getting to know your data, in particular the implications of missing values, strategies for filtering and interpolation, and how we use variance in EDA. We will also dig further into how we identify covariance.

Essential Readings

– **Wickham, Hadley and Garrett Grolemund (2017).** *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* **1st edition.** https://r4ds.had.co.nz/. **O'Reilly Media.**

Read Sections 7.1-7.4.

Review of summary statistics from Wickham, strongly recommended as prep.

– **Wickham, Hadley and Garrett Grolemund (2017).** *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data.* **1st edition.** https://r4ds.had.co.nz/. **O'Reilly Media.**

Read Sections 7.4-7.8.

– **Bruce, Peter, Andrew Bruce, and Peter Gedeck (2020).** *Practical Statistics for Data Scientists, 2nd Edition.* **Second.** [https://learning.oreilly.com/library/view/practical-statistics-for/9781492072935/.](https://learning.oreilly.com/library/view/practical-statistics-for/9781492072935/) **O'Reilly Media.**

Read Chapter 1: Exploratory Data Analysis.

Class #07: Wednesday 13 October 2021

Reading Week

Students will use this week to work on their detailed outlines and exploratory data analyses.

Class #08: Wednesday 20 October 2021

Exploratory Data Analysis II

In this second half of the EDA module we will focus on visualization, in particular aesthetics that help intuitively convey baseline trends, missing values, covariance, and introduce more sophisticated visualizations such as mapping, and how the aesthetics we have used thus far facilitate visualizations of cluster methods we will be using later in the course.

Part III: Visualizing Trends

Class #09: Wednesday 27 October 2021

Mapping

For data with a geographic component, maps are one of the most compeling ways to effectively communicate with your audience. Here, we ask (and answer!): where (geographically or spatially) are we seeing trends and how does it vary across these spaces? In the first part of the class we will use the grammar of graphics to create maps of regions we are interested in, then overlay the results of our EDA(s) onto those maps. In the latter half of the class we will explore how different aesthetics can be used to highlight different facets of our data and analyses, in particular the potential to overlay too much data onto a map, undermining the communication effort.

Essential Readings

– **Healy, Kieran (2018).** *Data Visualization: A Practical Introduction.* **1st edition.** [https://socviz.co/.](https://socviz.co/) **Princeton, NJ: Princeton University Press.**

Read Chapter 7.

We will use the Healy reading to explore a simple form of mapping to *understand the concepts*. The remaining readings illustrate the libraries we will be using in the learning exercises and your projects: the family of packages based on the spatial features package `sf`.

– **Moreno, Mel and Mathieu Basille (2018).** *Drawing Beautiful Maps Programmatically with R, Sf and Ggplot2 — Part 1: Basics.* [https://www.r-spatial.org/r/2018/10/25/ggplot2-sf.html.](https://www.r-spatial.org/r/2018/10/25/ggplot2-sf.html)

– **Moreno, Mel and Mathieu Basille (2018).** *Drawing Beautiful Maps Programmatically with R, Sf and Ggplot2 — Part 2: Layers.* [https://www.r-spatial.org/r/2018/10/25/ggplot2-sf-2.html.](https://www.r-spatial.org/r/2018/10/25/ggplot2-sf-2.html)

– **Moreno, Mel and Mathieu Basille (2018).** *Drawing Beautiful Maps Programmatically with R, Sf and Ggplot2 — Part 3: Layouts.* [https://www.r-spatial.org/r/2018/10/25/ggplot2-sf-3.html.](https://www.r-spatial.org/r/2018/10/25/ggplot2-sf-3.html)

Optional Readings

– **Kahle, David and Hadley Wickham (2013). "Ggmap: Spatial Visualization with Gg-plot2''. In:** ***The R Journal*** **5.1.** https://journal.r-project.org/archive/2013/RJ-2013-014/index.html**, pp. 144–161.**

  A bit deeper a dive into `ggmaps` for those interested in mapping for their final project.

– **Lansley, Guy and James Cheshire (2016).** ***An Introduction to Spatial Data Analysis and Visualisation in R.*** **London, UK: University College London, p. 121.**

  A bit technical, but worth looking at the later chapters for illustration of spatial autoregressions and inspirations on what to do and what not to do with map visualizations.

Class #10: Wednesday 03 November 2021

Cluster Analysis I

Descriptive statistics are great for looking trends in individual variables and whether *small* subsets of variables move together or not, but we often want to see if we can identify discernable groups, or *clusters*, in a larger dataset comprising many variables. In this class we will learn how to use two of the most common clustering strategies: hierarchical clustering and k-means. In the first half of the class we will visually explore how these clustering strategies work, in effect how they "think about" which observations are similar and which are not. As we will see, sometimes clusters give us compeling, intuitive results; in other cases, they give us gibberish. In the second half of the class we will discuss the role expert qualitative knowledge necessary to interepret cluster analyses and how this contributes to effective, credible hypothesis generation.

Essential Readings

– **James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2017).** ***An Intro-duction to Statistical Learning: With Applications in R.*** **Springer Texts in Statistics 103. New York: Springer.**

  Read Introduction to Chapter 10, Sections 10.1, then 10.3.

  This is the most "mathy" text you will see in the readings; read these introductions to get a feel for how unsupervised learning, in particular clustering, differs from predictive models (regression) many of you are already familiar with. You should read the following items from Imai and Peng, then come back to the remainder of these sections on k-means and heirarchical clustering.

– **Imai, Kosuke (2018).** ***Quantitative Social Science: An Introduction.*** **Princeton: Princeton University Press.**

  Read pages 111–115.

  This is one fairly accessible introduction to k-means clustering. After we do a couple of toy clustering examples, we will then explore the congress data Imai uses in this discussion.

– **Peng, Roger D. (2016).** ***Exploratory Data Analysis with R.*** https://bookdown.org/rdpeng/exdata/**.**

  Read Section 12 K-means Clustering.

– **Peng, Roger D. (2016).** ***Exploratory Data Analysis with R.*** https://bookdown.org/rdpeng/exdata/**.**

  Read Section 11 Heirarchical Clustering.

– **James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2017).** ***An Intro-duction to Statistical Learning: With Applications in R.*** **Springer Texts in Statistics 103. New York: Springer.**

Read Sections 10.3.1–10.3.3.

Read these sections to confirm your understanding from the presentations of the algorithms in Imai and Peng. Do not spend an inordinate amount of time on the math. Pay close attention to the discussion in Section 10.3.3 on validation, interpretation, and clustering strategies.

Class #11: Wednesday 10 November 2021

Cluster Analysis II

We will continue to explore how variables and indvidual observations contribute to clusters. We will then use variants of principal components analysis (PCA), as well as a few other strategies for helping us understand whether the clusters generated by our clustering algorithms are meaningful. We will conclude by illustrating strategies for comparing different sets of clusters as a way to determine which is most effective for the trends and objectives of our analyses.

Essential Readings

– **Kassambara, Alboukadel (2017).** *Principal Component Analysis Essentials.* http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/112-pca-principal-component-analysis-essentials/.

Kassambara provides an accessible introduction and tools for principal component analyses. A number of the examples in the lectures are based on Kassambara's examples, so it would be beneficial to work through them before class.

– **Kassambara, Alboukadel (2017).** *CA - Correspondence Analysis in R: Essentials - Articles - STHDA.* http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/113-ca-correspondence-analysis-in-r-essentials/.

This reading introduces correspondence analysis, essentially PCA for categorical data.

– **Kassambara, Alboukadel (2017).** *FAMD - Factor Analysis of Mixed Data in R: Essentials - Articles - STHDA.* http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/115-famd-factor-analysis-of-mixed-data-in-r-essentials/.

Finally, the FAMD methods described in this reading illustrate how to perform principal component analyses on mixed data sets, those with both quantitative and categorical variables. For many of you doing clustering analysis, this will be the set of tools you will be using.

Optional Readings

– **James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2017).** *An Introduction to Statistical Learning: With Applications in R.* **Springer Texts in Statistics 103. New York: Springer.**

Skim Section 10.2.

This section of James et. al. provides a bit more of the math behind principal component analyses. It is useful to skim for the intuition, but it is not necessary for the EDA we will be doing in this course.

Class #12: Wednesday 17 November 2021

Text Mining I

Not all data is easily or readily available as quantitative, or even structured, data sets. Identifying trends in groups of text documents (a corpus) is often a powerful complement to traditional qualitative analysis. In this class we will introduce the principles of text mining, its strengths and limits, and the fundamentals of transforming a document into data structures we can use to identify trends. Next we will explore a rather intuitive form of text mining: sentiment analysis. Finally, we will begin our first discussion on the strengths,

weaknesses, and potential pitfalls in text mining, how how to navigate these in the process of hypothesis generation.

Essential Readings

– **Silge, Julia and David Robinson (2017).** *Text Mining with R: A Tidy Approach.* **1st edition. https://www.tidytextmining.com/. O'Reilly Media.**

Read Chapters 1 and 2.

Class #13: Wednesday 24 November 2021

Thanksgiving

Go eat turkey. Tell your family how much you love data science :)

Class #14: Wednesday 01 December 2021

Text Mining II

In this lecture we will continue our exploration of text mining by working through two common analyses: frequency analysis and relationships amongst groups of words. We will also see yet another instance of integrating analyses, representing relationships between words as networks. In the second half of the lecture we will continue our focus on visualization and how this contributes to hypothesis generation.

Essential Readings

– **Silge, Julia and David Robinson (2017).** *Text Mining with R: A Tidy Approach.* **1st edition. https://www.tidytextmining.com/. O'Reilly Media.**

Read Chapters 3 and 4.

Class #15: Wednesday 08 December 2021

Project Presentations

In this class students will present their final projects.

Assignments and Writing Guidelines

The following provides details on the assignments for the course. All written assignments *must be submitted in PDF format* and conform to the writing format guidelines. Policy research project milestones and learning exercises will be submitted via GitHub. **All assignments are due *at or before 2359* on the assignment due date.**

Assignments

The best way to learn data science is by doing. In addition to the policy research project, you will also have 5 learning exercises.

As a note on writing requirements, the length requirements for assignments and project milestones are described in terms of *maximum* word count. For the page equivalent, 500 words is *approximately* 1 page with 1 inch margins, **single-spaced**, 12 pt Times New Roman (or a very similar serifed font like the one used in this document).

Participation

You will start with 7 (out of a possible 10) points. You are expected to have completed all of the readings for a given week before the class and be ready to discuss those readings in class. Portions of each class will be run seminar style, with the expectation that students discuss the concepts and issues at hand in a civil, constructive, yet rigorously analytic manner. This is your opportunity to gain points. If you attempt to participate and clearly demonstrate you have not done the related reading, you will definitely lose participation points. That said, informed, thoughtful, civil, and constructive disagreement with other students or Dr. Sowell is encouraged, especially when Dr. Sowell makes intentionally leading, biased, or contradictory assertions to encourage discussion and creativity.

Also, as a note on participation, we will be using laptops extensively for exercises in the course. Beyond this, Dr. Sowell encourages students to *quickly* look up relevant materials online to contribute during discussion. That said, refrain from spending class time on e-mail, social media, instant messaging, or anything else that is not directly related to the class or discussion at hand.

Learning Exercises

The learning exercises are intended to be short sets of exercises to further familiarize yourself with topics from class. These will be distributed via the course GitHub repository. They will be submitted by pushing the completed files to your personal GitHub repository for the course. We will set up personal repositories during the second class meeting on Wednesday 08 September 2021.

Policy Research Project

The objective of the policy research project is for you to apply lessons and concepts learned in the course to a policy or governance issue area *of your choice*. For the policy research project you will identify a data set (or sets) that you would like to explore, visualize, and evaluate.

Your first task is to Identify a data set of interest to you, that you believe you can do an in-depth analysis of for your policy research project. Your proposal should demonstrate you are generally familiar with the data and you have done some basic analysis. Dr. Sowell will have to approve this data set. You should select your data set early and set up a time to discuss the data set with Dr. Sowell **well before** the proposal deadline to get preliminary approval.

There is no minimum word limit on the policy research research project. The maximum word limit is 10,000 words, approximately 20 pages, not including figures or references. See the writing guidelines for specifics on what does and does not contribute to the word count.

Students are *required* to maintain the references used in their assignments, in particular for the policy research project, in Zotero. If the student has not already, the student should create a Zotero account.

Policy research project milestones:

1. **Zotero Policy Research Project Group:** due *Friday 03 September 2021*

   Deliverables:

   – e-mail to Dr. Sowell (jsowell@tamu.edu) confirming the student has
     1. installed and setup Zotero
     2. joined the Zotero group for this course
     3. created their own group for this course

   Students will use Zotero to create a shared group (library) entitled `Z - DS - lastname` (formatted exactly as here, with the spaces, where `lastname` is your last name) and use Zotero's group invitation function to invite Dr. Sowell (jsowell@tamu.edu) to the group. The Zotero Group should be private and Dr. Sowell should have edit rights so he can share references with the student. **This task is worth 3 points of your final grade. If you do not set up and share your Zotero library by 2359 on you will lose these points permanently.** For step-by-step instructions on setting up

Zotero, see the section on Zotero Configuration in the Appendix. If you have any questions or run into any problems, please e-mail Dr. Sowell (jsowell@tamu.edu) at least one hour before class on .

All references used in the policy research project *must* be saved in the student's shared Zotero library. This will make your life a lot easier: you can easily copy references from the course library to your library for use in your policy research project, it allows Dr. Sowell to review your policy research project references with you, and allows Dr. Sowell to share relevant references with you when appropriate. When adding references to journal articles, reports, etc., you should make sure the PDF of the document is attached to that entry. The Zotero Connector will often do this for you, but you should double check and add the PDF if it does not.

Students will maintain data and R project files for milestone deliverables in their GitHub repository. We will setup the GitHub repositories in the second class meeting on Wednesday 08 September 2021.

2. **Proposal:** *(5%)* due *Monday 20 September 2021*

   Deliverables:

   – **.Rmd file** of the proposal document (in the `templates` directory of your final project repository), with proper references and citations to the dataset to be used
   – **PDF** of the proposal document
   – data should be in the `data` folder of your project repository
   – **.Rmd** document (in the `templates` directory of you final project repository) illustrating baseline data import and simple summary statistics of your dataset(s), including variables of interest, their types (number, categorical, text), and a brief description of each variable.

   The proposal should be a 500 word (max) description of the policy or governance issue you plan to address with the dataset(s) you have selected for your policy research project.

   At a minimum, the proposal document should reference readings from the course (with a bibliography that does not count against the 500 word limit). A *good* proposal will also include references to materials outside the course that (1) support the analysis strategies and arguments in the proposal and (2) demonstrates the student has already started their own research on the topic.

   In terms of the structure of the proposal, it should answer the following questions:

   1. What is the issue you are addressing?
   2. Why is this issue important?
   3. What data are you using to explore this issue?
   4. What has already been done with this data?
   5. What questions are you trying to answer? Relate your questions to what has already been done, this will require some amount of literature review relative to your dataset.
   6. Why is what you are doing different from other analyses and visualizations you have found? How does this data exploration contribute to answering your questions (again relative to other work)?

3. **Detailed Outline and EDA:** *(20%)*, due *Monday 15 November 2021*

   Deliverables:

   – .Rmd file of the detailed outline document with the EDA appendix
   – PDF of the detailed outline

   The objective of the detailed outline is to articulate the fundamental structure of the project report, the current status of analyses and arguments, supporting materials, and how these are used to support the analysis and argument. The outline is intended to get the student thinking about the structure of the argument; it is expected to change based on feedback and further analysis leading up to the rough draft. That said, the detailed outline should articulate a clear and coherent narrative, argument, and supporting analysis.

   You should copy the template for the final policy research project in the `templates` directory of your project GitHub repository as the template for your detailed outline. The outline portion should flesh

out the sections of the body of the policy research project following the guidelines below. The detailed instructions for the EDA, which should be in the appendix of your detailed outline, is in the EDA section (in the appendix) of the policy research project template file.

The outline (the body of the document) should comprise

- – a title (title, name, date, no abstract at this point)
- – the approved proposal, followed by
- – the outline with enumerated headers (1, 2.3, 5.6.3, etc.) for the major sections, subsections, etc.
- – enumerated sections should include an "opener" and a "closer" that conveys the content for that section:
  - – the overall objective of the detailed outline is to tell the high-level story of the argument and analysis
  - – the "opener" (1-3 sentences) is like an opening paragraph: it introduces the topic, problem, argument, or analysis to be presented in that section
  - – the "closer" (1-3 sentences) is like the concluding paragraph: it articulates the take-aways of a narrative, summary of the problem, highlights of the argument, or conclusions of an analysis
  - – sentences in both the opener and closer should be substantive declaratives, **not** "This section will do this" or "This section will show that"
- – the exploratory data analysis (EDA, the appendix) is an opportunity for you to demonstrate you have systematically explored your data set; the details of the EDA specification can be found in the appendix of the policy research project template document in your GitHub repository

4. **In Class Project Presentation:** *(5%)*, due *Tuesday 07 December 2021*

   Deliverables:

   - – PDF of the presentation slides

   On the last day of class students will present their policy research projects. Each student presentation will be approximately 20 minutes followed by discussion. Slides are required for the presentation. Slides should uploaded to the `Project Presentation` collection in the student's Zotero library at least two hours before class on .

5. **Policy Research Project Report:**, *(30%)* due the last day of exam week *Wednesday 08 December 2021*

   Deliverables:

   - – .Rmd of the final policy research project report
   - – PDF of the final policy research project report

   Final policy research project reports will follow the same format as the rough draft. The final project report should incorporate changes, suggestions, and comments made on the rough draft. Dr. Sowell will compare the rough draft and final draft when grading the final report.

Write-Up Formatting Guidelines

The following describes Dr. Sowell's standard writing guidelines. That said, we will be using an R Markdown template for all of our assignments and your policy research project. While these requirements may *seem* daunting, the RMarkdown template will take care of most of them. Following the ethos of Markdown and the reproducible research paradigm, the objective is to get the formatting out of the way so the analyst can focus on writing and analysis.

These guidelines are not optional and will be strictly enforced. If you submit material that does not conform to these guidelines, it will be returned ungraded and with a 10% late penalty.

- – single-spaced
- – title page with title, name of author, and abstract; title page should not have a page number

- – title page and abstract (approximately 300 words) do not contribute to word count for assignment unless otherwise specified
- – font should be New Times Roman or similar serifed font
- – font size for abstract and body of text should be 12 pt
- – document should be fully justified as in books and journal articles, no ragged right edge
- – use enumerated footnotes, 10 pt; *do not ever use endnotes*
- – 1 inch margins all around (left, right, top, bottom, this is standard in Word)
- – block quotes consistently inset from left and right margins
- – page enumeration in footer, no page number on title page, body enumeration starts starts at page 2
- – enumerate sections and subsections (1, 2.1, 3.5.2, etc.)
- – figures should be labeled ("Figure 1: Scatter plot of data set X", "Figure 2: Distribution of variables in category Y", etc.), referenced by figure number ("Note that the distribution in Figure 4 is left skewed. . . "); figure labels will contribute to word count
- – references must be stored in Zotero
- – inline references and bibliography should follow Chicago author-date format
- – the bibliography will not contribute to the word count
- – inline references:
    - – materials (articles, books, etc.) with page numbers must include the page number or page range that includes the quote or evidence referenced
    - – materials, such as web pages that are not enumerated, should include the finest grained subsection containing the quote or evidence where the page number or page range would be in the inline reference
    - – inline references that do not follow these guidelines will result in assignment returned with a 10% penalty
- – documents submitted should be in PDF format and should allow highlighting of text using PDF annotation tools such as Adobe Acrobat Reader; you should check this as you write and before submitting, exotic invisible formatting in Word occasionally breaks this requirement
- – PDF documents will be submitted electronically either via Zotero or via e-mail to jsowell@tamu.edu depending on the assignment guidelines; hard copy will not be accepted

Dr. Sowell will provide an example PDF to illustrate these guidelines. When grading your assignments, Dr. Sowell will annotate your document electronically. Any mainstream PDF reader, such as Adobe Acrobat Reader, Skim, or Apple's Preview will render these comments. A comment attached to the title of the assignment will contain the total grade and overall comments. Graded essays and write-ups will be returned to you electronically.

University Policies

Attendance

The university views class attendance and participation as an individual student responsibility. Students are expected to attend class and to complete all assignments.

Please refer to Student Rule 7 in its entirety for information about excused absences, including definitions, and related documentation and timelines.

Other absences may be excused at the discretion of the instructor with prior notification and proper documentation. In cases where prior notification is not feasible (e.g., accident or emergency) the student must provide notification by the end of the second working day after the absence, including an explanation of why notice could not be sent prior to the class.

On some occasions, the instructor may have to miss a class due to administrative or academic responsibilities out of town. If it does occur, the instructor reserves the right to reschedule class at a time when the vast majority of students are available for the make-up class and will convey the material to students unable to attend the make-up during office hours.

Makeup Work Policy

Students will be excused from attending class on the day of a graded activity or when attendance contributes to a student's grade, for the reasons stated in Student Rule 7, or other reason deemed appropriate by the instructor.

Please refer to Student Rule 7 in its entirety for information about makeup work, including definitions, and related documentation and timelines.

"Absences related to Title IX of the Education Amendments of 1972 may necessitate a period of more than 30 days for make-up work, and the timeframe for make-up work should be agreed upon by the student and instructor" (Student Rule 7, Section 7.4.1).

"The instructor is under no obligation to provide an opportunity for the student to make up work missed because of an unexcused absence" (Student Rule 7, Section 7.4.2).

Students who request an excused absence are expected to uphold the Aggie Honor Code and Student Conduct Code (see Student Rule 24).

Academic Integrity Statement and Policy

"An Aggie does not lie, cheat or steal or tolerate those who do."

"Texas A&M University students are responsible for authenticating all work submitted to an instructor. If asked, students must be able to produce proof that the item submitted is indeed the work of that student. Students must keep appropriate records at all times. The inability to authenticate one's work, should the instructor request it, may be sufficient grounds to initiate an academic misconduct case" (Section 20.1.2.3, Student Rule 20).

You can learn more about the Aggie Honor System Office Rules and Procedures, academic integrity, and your rights and responsibilities at aggiehonor.tamu.edu.

Dr. Sowell strongly encourages reading groups for discussing course materials, but not for distributing the reading load. Dr. Sowell also recognizes the role and efficacy of group learning and peer review of assignment deliverables, such as proof-reading one another's work and/or discussing the structure and flow of arguments presented in assignments. If you engage in this kind of collaboration, you must add a footnote to the your name (as the author) with a statement indicating who you collaborated with and how they contributed to the work you are turning in under your name. As an example, "John Smith proof-read a draft of this assignment, providing editorial comments and suggesting I rearrange the order of my cases to improve the logical flow of my case studies section." Another example would be "I discussed this assignment with Jane Smith and she suggested the articles (Warner 2016; Billings 1967), which I have included in this work."

Americans with Disabilities Act (ADA) Policy

Texas A&M University is committed to providing equitable access to learning opportunities for all students. If you experience barriers to your education due to a disability or think you may have a disability, please contact Disability Resources in the Student Services Building or at (979) 845-1637 or visit disability.tamu.edu. Disabilities may include, but are not limited to attentional, learning, mental health, sensory, physical, or chronic health conditions. All students are encouraged to discuss their disability related needs with Disability Resources and their instructors as soon as possible.

Title IX and Statements on Limits to Confidentiality

Texas A&M University is committed to fostering a learning environment that is safe and productive for all. University policies and federal and state laws prohibit gender-based discrimination and sexual harassment, including sexual assault, sexual exploitation, domestic violence, dating violence, and stalking.

With the exception of some medical and mental health providers, all university employees (including full and part-time faculty, staff, paid graduate assistants, student workers, etc.) are Mandatory Reporters and

must report to the Title IX Office if the employee experiences, observes, or becomes aware of an incident that meets the following conditions (see University Rule 08.01.01.M1):

– The incident is reasonably believed to be discrimination or harassment.
– The incident is alleged to have been committed by or against a person who, at the time of the incident, was (1) a student enrolled at the University or (2) an employee of the University.

Mandatory Reporters must file a report regardless of how the information comes to their attention – including but not limited to face-to-face conversations, a written class assignment or paper, class discussion, email, text, or social media post. Although Mandatory Reporters must file a report, in most instances, you will be able to control how the report is handled, including whether or not to pursue a formal investigation. The University's goal is to make sure you are aware of the range of options available to you and to ensure access to the resources you need.

Students wishing to discuss concerns in a confidential setting are encouraged to make an appointment with Counseling and Psychological Services (CAPS).

Students can learn more about filing a report, accessing supportive resources, and navigating the Title IX investigation and resolution process on the University's Title IX webpage.

Statement on Mental Health and Wellness

Texas A&M University recognizes that mental health and wellness are critical factors that influence a student's academic success and overall wellbeing. Students are encouraged to engage in proper self-care by utilizing the resources and services available from Counseling & Psychological Services (CAPS). Students who need someone to talk to can call the TAMU Helpline (979-845-2700) from 4:00 p.m. to 8:00 a.m. weekdays and 24 hours on weekends. 24-hour emergency help is also available through the National Suicide Prevention Hotline (800-273-8255) or at suicidepreventionlifeline.org.

Appendix

Zotero Configuration

The following instructions describe how to set up the Zotero App and create a shared group library.

Zotero App Setup

1. Create a Zotero account at https://www.zotero.org using your `@tamu.edu` e-mail address.

2. Install

    1. Zotero app, available at https://www.zotero.org/download/
    2. Install the Zotero Connector web browser plugin, available at https://www.zotero.org/download/connectors

3. You should receive an invitation to the course library in both the e-mail you set up your account with and in the Zotero Inbox, available via the Zotero web interface.

4. To confirm your Zotero app is syncing with the course library, you should check your Zotero app preferences. In the `Preferences` window, select the `Sync` tab and confirm that

    – Zotero shows `Username: your_username` (where `your_username` is your username)
    – you have checked `Sync Automatically` and `Sync full-text content`
    – you have checked `Sync attachment files in My Library using Zotero` and selected `Download files at sync time`
    – you have checked `Sync attachment files in gorup libraries using Zotero storage` and selected `Download files at sync time`

    A screenshot illustrating what your settings should look like can be found below.
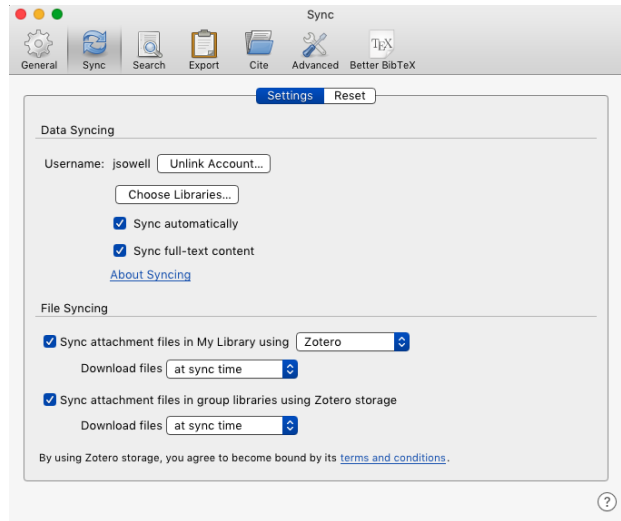
If you prefer not to sync automatically, uncheck `Sync automatically`. ***If you choose this option you will have to explicitly sync your libraries using the small circular green arrow in the upper right of the Zotero app.***

Shared Groups (Library) Setup

1. Log in to the Zotero web interface at https://www.zotero.org

2. Click on the `Groups` link

3. Click `Create a New Group` link directly under the header `Zotero Groups`

4. Add a name for your shared library where it says `Choose a name for your group` and select the `Group Type` as `Private Membership` as illustrated below
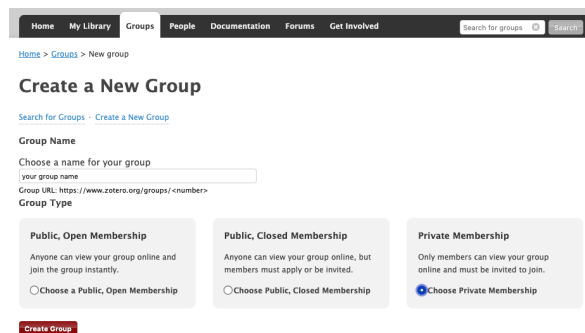
5. Click the `Create Group` button

6. Select the following group settings (illustrated in the screenshot below):

   – for `Group Type`, select `Private`
   – for `Library Reading`, select `Any group member`
   – for `Library Editing`, select `Any group members`
   – for `File Editing`, select `Any gorup members`

   then click `Save Settings`; these are the defaults, so you should not have to change anything.

7. To add new members to the group, click the blue link `Member Settings` link under the heading `groupname: Member Settings` and click the link `Send More Invitations` at the bottom of the page and follow the instructions there.

Notes on Zotero Connector Plugin

The Zotero Connector adds a small icon to the right of the address bar in your web browser (upper right corner of the window). To use the Connector, the Zotero app must be open. By default, when you click the Zotero icon to download a given reference, it will automatically put that reference *in whichever folder you currently have selected in the Zotero app.* This is quite convenient if you have organized your research folder into topic specific subfolders, or, in my case, if you have it organized by class and category of reading material (essential, optional, etc.).